

The Evolutionary Ebb and Flow
of
Genomic Nucleotide Content

Hamid Nikbakht

Thesis In the Department of Biology

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Biology) at

Concordia University

Montreal, Quebec, Canada

December 2012

© Hamid Nikbakht, 2012

**CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Hamid Nikbakht

Entitled: The Evolutionary Ebb and Flow of Genomic Nucleotide Content

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Biology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

<u>Dr. Judith Kornblatt</u>	Chair
<u>Dr. Xuhua Xia</u>	External Examiner
<u>Dr. Xiaowen Zhou</u>	External to Program
<u>Dr. Robert Weladji</u>	Examiner
<u>Dr. Vincent Martin</u>	Examiner
<u>Dr. Donal Hickey</u>	Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

Dean of Faculty

Abstract

Nucleotide content is one of the most obvious and most easily quantifiable features of a genome. Consequently, all reports of new genome sequences include a report on the nucleotide content, i.e., the frequencies of each of the four nucleotide bases in the genome sequence. Despite the ease with which nucleotide content can be measured, however, it has proved to be much more complicated to explain the evolution of the genomic nucleotide content. A number of different explanations have been proposed. These include the effects of biased mutational patterns (i.e. neutralist models) such as biased DNA repair during recombination, as well as the effect of natural selection on shaping the genomic nucleotide content (i.e. selectionist models). A well known example for the latter group of theories is the effect of temperature in selecting for higher GC content to enhance the thermostability of the DNA. However, many of these theories do not address the problem of variations in nucleotide content between different regions of a single genome, i.e., intragenomic compositional heterogeneity. In this study, I investigate the evolutionary behaviour of the genomic nucleotide content in a group of model organisms and show that the current state of the nucleotide content of a genome by itself cannot always explain the entire evolutionary history of that genome and, in fact, the sequence of events happening to a genome regarding its nucleotide content can be much more complicated than how it looks. For example I show that having an unbiased content doesn't necessary mean a stationary nucleotide substitution model in an organism. In this study I show that the bias in nucleotide content of a genome can change its direction several times and this going back and forth can actually happen very fast; fast enough to trace the results of this ebb and flow within the same genus. I present an example of this evolutionary ebb and flow of genomic nucleotide content within genus *Plasmodium*. I also discuss that these exceptional behaviours of the nucleotide contents of the genomes can, indeed, affect other features of living organisms such as the substitution rates

between different nucleotides as well as their protein content, and consequently they can interfere with phylogenetic studies. In order to explain the heterogeneities within a single genome, I investigate the relationship between gene length and the degree of nucleotide bias, and find that these two parameters show a significant negative correlation. This lead me to propose a mechanism that can explain heterogeneity in the nucleotide content of the genome, without having to invoke variations in mutational patterns between different parts of the same genome. My proposed model resembles Charlesworth's "Background Selection" model. My findings shed light on the importance of the evolutionary behaviour of the genomic nucleotide content to be considered in studying different features of an organism as well as studying the evolutionary relationship between the organism of our interest and other related species.

Table of Contents

Abstract	III
Table of Contents.....	V
List of Figures	VII
List of Tables	X
List of Abbreviations	XI
Chapter 1. Introduction and Literature Review	1
1.1. Measuring Nucleotide Content	1
1.2. Variation in Nucleotide Content between Genomes	4
1.3. Variation in Nucleotide Content within Genomes	5
1.4. Explaining the Variations in Nucleotide Content between and within Genomes	7
1.5. Phylogenetic inference is affected by nucleotide bias	14
1.6. Thesis hypothesis and goals.....	23
Chapter 2. Rapid reversal of nucleotide bias within the genus Plasmodium	27
2.1. Introduction.....	27
2.2. Methods.....	30
2.3. Results	34
2.4. Discussion	45
Chapter 3. Nucleotide bias can distort the usual transition-transversion ratio	48
3.1. Introduction.....	48
3.2. Methods.....	51
3.3. Results:.....	56
3.4. Discussion:	76
Chapter 4. The relationship between gene length and nucleotide content	79
4.1. Introduction.....	79
4.2. Methods.....	83
4.3. Results	86
4.4. Discussion	99
Chapter 5. General Discussion	101
5.1. The ebb and flow of genomic nucleotide content	101
5.2. Distortion of the usual transition-transversion ratio.....	109

5.3. The relationship between gene length and nucleotide content	128
5.4. Future Research Directions.....	137
References	140
Appendices.....	151

List of Figures

Figure 1.1: Schematic representation of different nucleotide substitutions between four different nucleotides	21
Figure 2.1: Distribution of GC content in coding sequences of two <i>Plasmodium</i> species	35
Figure 2.2: The distribution plot of the GC content of 4283 pairs of orthologous genes between the two <i>Plasmodium</i> species	36
Figure 2.3 A: GC content distribution of the orthologous pairs in the two <i>Plasmodium</i> species plotted against their alignment length.....	37
Figure 2.3 B: GC content distribution of the 14 lowest GC content <i>P. vivax</i> coding sequences (GC contents lower than 30%) and their <i>P. falciparum</i> orthologous sequences plotted against their alignment length.....	37
Figure 2.4: GC content of the dataset of orthologous coding sequences in <i>P. falciparum</i> and <i>P. vivax</i>	42
Figure 2.5 (A, B, and C): GC content of the dataset of orthologous coding sequences in <i>P. falciparum</i> and <i>P. vivax</i> in three codon positions	43
Figure 2.6: Individual nucleotide content of the dataset of orthologous coding sequences in <i>P. falciparum</i> and <i>P. vivax</i>	44
Figure 3.1: Schematic representation of different nucleotide substitutions between four different nucleotides	49
Figure 3.2: Nucleotide substitution profile between the mitochondrial genomes of two <i>Plasmodium</i> species; <i>P. falciparum</i> and <i>P. vivax</i>	56
Figure 3.3: Nucleotide substitution profile between mitochondrial genomes of <i>P. falciparum</i> and <i>P. reichenowi</i>	66
Figure 3.4: Nucleotide substitution profile between two datasets of three concatenated, aligned orthologous mitochondrial genes in human and chimpanzee.....	69
Figure 3.5: Nucleotide substitution profile between two datasets of six concatenated, aligned orthologous nuclear genes in <i>P. falciparum</i> and <i>P. vivax</i>	71
Figure 3.6: Nucleotide substitution profile between two datasets of five concatenated, aligned orthologous nuclear genes in human and chimpanzee.....	73

Figure 3.7: The observed pattern of transitional changes between two nucleotides from the same class, between two sets of concatenated, aligned orthologous nuclear genes in <i>P. falciparum</i> and <i>P. vivax</i>	75
Figure 4.1: GC content distribution in honeybee genome	86
Figure 4.2: Average length of coding sequences in two groups of honeybee genes with lower and higher GC contents.....	88
Figure 4.3: The relationship between GC content of the coding sequences with average length of (A) coding sequences (B) total gene sequences and (C) introns in honeybee	89
Figure 4.4: The changes in GC content of each codon positions compared to the changes in the GC content of the coding sequences in honeybee.....	91
Figure 4.5: Proportion of different groups of amino acids vs. GC content of their corresponding coding sequences in honeybee.....	92
Figure 4.6: The distribution of GC content in honeybee and rice coding sequences	94
Figure 4.7: Average length of coding sequences in two groups of genes in honeybee and rice...	96
Figure 4.8: GC content distribution in two subsets of coding sequences in honeybee	97
Figure 5.1: Schematic representation of the nucleotide substitutions between mitochondrial genome of <i>P. falciparum</i> and <i>P. vivax</i>	111
Figure 5.2: Schematic representation of the nucleotide substitutions between mitochondrial genome of <i>P. falciparum</i> and <i>P. reichenowi</i>	116
Figure 5.3: The relationship between the percentage sequence differences and the divergence time based on Jukes-Cantor model in slow (dotted line) and fast (dashed line) evolving sites..	117
Figure 5.4: Schematic representation of nucleotide substitutions between two sets of concatenated, aligned orthologous mitochondrial genes in human and chimpanzee	121
Figure 5.5: Schematic representation of nucleotide substitutions between two sets of six concatenated, aligned orthologous nuclear genes in <i>P. falciparum</i> and <i>P. vivax</i>	122
Figure 5.6: Schematic representation of nucleotide substitutions between two sets of concatenated, aligned orthologous nuclear genes in human and chimpanzee	123
Figure 5.7: Schematic representation of nucleotide substitutions between two nucleotides from the same class, between two sets of concatenated aligned orthologous nuclear genes in <i>P. falciparum</i> and <i>P. vivax</i>	125

Figure 5.8.A: Nucleotide substitution pattern between nuclear genomes of human and chimpanzee.....	126
Figure 5.8.B: Nucleotide substitution pattern between mitochondrial genomes of <i>P. falciparum</i> and <i>P. vivax</i>	127
Figure 5.9: Average length of coding sequences in two groups of genes in honeybee and rice.	131
Figure 5.10: The negative relationship between bias in nucleotide content and average length in genes.....	132

List of Tables

Table 2.1: GC content of seven <i>Plasmodium</i> species	34
Table 2.2: A part of the table containing the data for GC content of the total set of orthologous coding sequences (4283 coding sequences) in two <i>Plasmodium</i> species	39
Table 2.3: The GC content and number of conserved sites between the two <i>Plasmodium</i> species.....	39
Table 2.4: The GC content and number of variable sites between the two <i>Plasmodium</i> species.	40
Table 2.5: The Average GC content in different sites of coding sequences in two <i>Plasmodium</i> species	41
Table 3.1: Nucleotide content in different sites of three mitochondrial genes in <i>P. falciparum</i> and <i>P. vivax</i>	58
Table 3.2: Nucleotide content of three mitochondrial genes in <i>P. falciparum</i> and <i>P. vivax</i>	61
Table 3.3: Nucleotide content of three mitochondrial genes in <i>P. falciparum</i> and <i>P. vivax</i> (in percentages)	62
Table 3.4: Nucleotide content of the variable sites in three mitochondrial genes in <i>P. falciparum</i> and <i>P. vivax</i>	63
Table 3.5: Nucleotide content of the variable sites in three mitochondrial genes in <i>P. falciparum</i> and <i>P. vivax</i> (in percentages)	64
Table 3.6: The <i>ts/tv</i> rate ratio between different genomes	77
Table 4.1: Average length of coding sequences in two groups of genes in honeybee and rice.....	95
Table 4.2: Number of coding sequences in two different datasets of honeybee coding sequences.....	98

List of Abbreviations

BLAST: Basic Local Alignment Search Tool

A: Adenine

C: Cytosine

G: Guanine

T: Thymine

GC Content: Percentage of Guanine + Cytosine Content

GC3: GC content at third codon position

ts: Transition

tv: Transversion

SCU: Selection on Codon Usage

Chapter 1. Introduction and Literature Review

1.1. Measuring Nucleotide Content

Deoxyribonucleic acid (DNA) is a polymer of four different mononucleotides, and the sequence of these four nucleotides encodes the genetic information needed for the proper functioning of the cell and the development of multicellular organisms. The four nucleotides Guanine (G), Cytosine (C), Adenine (A) and Thymine (T) are the building blocks of DNA. These four nucleotides can be divided into two groups based on their chemical structure. Adenine and Guanine are grouped as Purine nucleotides and Cytosine and Thymine are grouped as Pyrimidine nucleotides. Both Purines and Pyrimidines are built of heterocyclic compounds. Pyrimidines consist of a single heterocyclic compound with two Nitrogen atoms on positions 1 and 3 and Purines consist of a pyrimidine compound fused to an imidazole ring (Nitrogen atoms located on positions 7 and 9).

DNA molecules usually exist as anti-parallel double helices with nucleotide base pairing between Guanine and Cytosine bases and between Adenine and Thymine bases (Watson and Crick, 1953). Thus, if we imagine a molecule of DNA as a twisted ladder, the steps of the ladder will be pairs of nucleotides bond together through hydrogen bonds. In each pair of nucleotides there is one Purine and one Pyrimidine. There are three hydrogen bonds between Guanine and Cytosine and two hydrogen bonds between Adenine and Thymine. Further on, I will explain that this is a property about DNA structure that can have some evolutionary effects. The antiparallel structure of the DNA and the specific pairing between nucleotides causes equality between the content of Guanine and Cytosine as well as between Adenine and Thymine. Although Chargaff (1950) had discovered this equality three years before the discovery of the double helix structure of DNA by Watson and Crick. This equality is known as Chargaff's First Parity Rule (PR I).

Later, in 1968 Chargaff and his colleagues reported an extended equilibrium between complementary nucleotide (i.e. between Guanine and Cytosine as well as between Adenine and Thymine) (Rudner et al., 1968) in a single strand of DNA. This extended rule is referred as Chargaff's Second Parity Rule (PR II) in literature. It has been shown that PR II applies to most of eukaryotic nuclear genomes as well as eubacteria and archaeabacteria considering large extent of genomic DNA (Nikolaou C. and Almirantis Y., 2006).

Considering Chargaff's parity rules (I and II) one simply can conclude that if we know the proportion of one nucleotide within a genome, we can calculate the frequency of the three other nucleotides as well. For example, if the frequency of Guanine is 0.2, then the frequency of Cytosine is also 0.2. Since the combined frequencies of G+C would equal 0.4 in this case, the total frequency of the two remaining nucleotides, Adenine and Thymine, must equal 0.6. However, since the frequency of Adenine also is equal to the frequency of Thymine, (PR II), their individual frequencies come to 0.3 and 0.3. Thus, one can say that the frequencies of the four nucleotides in a genome have only one degree of freedom. This explains why the nucleotide content of a genome can be described by a single number, usually the GC content. This content can be different between different genomes of different species. Also as I will explain later in more details, different parts of a single genome can also show different GC content values. In Sections 1.2 and 1.3 I will talk more about the variation in the nucleotide content values between different genomes as well as within a single genome respectively.

The GC content of a genome can be calculated quite simply by counting the number of bases in the genome sequence using the following formula and is usually represented in percentages:

$$\text{GC Content} = \frac{G + C}{G + C + A + T} \times 100$$

For those genomes that have not been sequenced yet, the GC content can be inferred by experimental approaches. One way is using DNA melting point characteristic and spectroscopy in *UV* (wavelength of 260 nm). The absorbance of the DNA molecule in this wavelength increases when the double stranded structure of the DNA separates. The nucleotide content of a DNA molecule affects the temperature at which this phenomenon occurs. More specifically, the higher the GC content is, the greater the number of hydrogen bonds between the two strands of the DNA will be and, consequently, the higher the melting temperature will be (Marmur and Doty 1959). Another experimental approach to calculate the GC content of a genome without the knowledge of the actual sequence is the use of a CsCl density gradient ultracentrifugation. The nucleotide composition of a DNA molecule defines its molecular weight. Using the CsCl gradient ultracentrifugation one can conclude the molecular weight of the genomic DNA based on the molecule's position in the CsCl density gradient at sedimentation equilibrium (Sueoka et al. 1959). Yet another experimental method is using the Flow Cytometry technique. In this method, one can obtain the percentage of AT pairs as well as GC pairs along with the genome size using the AT-specific fluorochrome Hoechst 33258 (HO) and the GC-specific fluorochrome Olivomycin (OM) (Vinogradov A. E. 1994).

The motivation for reporting the GC content of a genome comes from the fact that the nucleotide content of the genomic DNA can affect different aspects of a species life and evolution. Aspects such as the protein content and the relative usage of synonymous codons (Shields et al., 1990; Gu et al., 1998, Singer and Hickey 2000) as well as the phylogenetic inference among different species (Foster and Hickey 1999) are shown to be affected by the genomic nucleotide content. The effect of the variation in the genome nucleotide contents on the phylogenetic inference can be seen in both DNA based phylogenetic studies as well as the amino

acid based studies. This will be discussed in Section 1.5 (i.e. Phylogenetic inference is affected by nucleotide bias) of this chapter.

1.2. Variation in Nucleotide Content between Genomes

As stated above, GC content can vary between the genomes of different species and also between different regions of a single genome. Inter-species variations cover a very wide range. For example, the genome of the malaria parasite *Plasmodium falciparum* has a very low GC content of only 19.44% (Gardner et al. 2002) whereas the GC content of the filamentous soil-dwelling Gram-positive bacterium *Streptomyces coelicolor* is very high at 72% (Bentley et al, 2002).

In general, closely related bacteria have been shown to represent similar GC contents. This led Lawrence and his colleagues (1997) to propose GC content to be considered as a reliable phylogenetic signal. Although later on, Gu and his colleagues (1998) showed that the GC content at synonymous codon positions – which are the most sensitive to nucleotide frequency changes – could not be used reliably to infer divergence times between species. In other words, even though apparently there is a general tendency for related species to show similar nucleotide contents in their genomes, there is not a positive, monotonic relationship between increasing phylogenetic distance and increasing differences in nucleotide content. Moreover, Foster and Hickey (1999) showed that, in fact, GC content can interfere with the correct phylogenetic inference because two species with close GC content values, regardless to their lack of biological relatedness, can group together. They showed that this false grouping can happen both when using DNA based phylogenetic inference methods as well as using protein based methods. They explained this by postulating that the genomic nucleotide content can affect the proteome content as well as the codon bias.

1.3. Variation in Nucleotide Content within Genomes

In addition to the large variations in the nucleotide content values between genomes, variations in nucleotide content have also been observed between different regions within a single genome. For instance, the nucleotide content of bacterial plasmids may differ from that of the main chromosome (van Passel et al. 2006). van Passel and his colleagues examined the sequences of 230 plasmids and their corresponding host chromosomes and found that the GC content (which they call it “genome signature” in their paper) of the plasmid sequences is “dissimilar” from that of their host genomes.

The most obvious examples of intragenomic variations in nucleotide content, however, can be seen among eukaryotic genomes, especially within large genomes such as the human genome. This within-genome heterogeneity in the nucleotide content can be seen by staining metaphase chromosomes with simple dyes such as Geimsa. Metaphase chromosomes stained with Geimsa following the digestion of chromosomal proteins by trypsin show a pattern of dark and light regions/bands. It is now known that the absorbance of the dye depends on the density of hydrogen bonding between the two DNA strands which, in turn, depends on the content of the Guanine and Cytosine pairs in an antiparallel structure of the DNA which corresponds with the genome’s GC content. Dark regions in G-Banding (Geimsa positive bands) are highly AT rich while light regions (Geimsa negative bands) are GC rich (Comings et al. 1973). Although we now understand the molecular basis of the staining pattern, such dyes have been used to identify the characteristic banding patterns of chromosomes long before the underlying causes of the differential staining was understood.

Although it was described for the first time in 1976 by Bernardi (Macaya et al. 1976), Bernardi and his colleagues coined the term “isochore” for the first time in 1981 (Cuny et al.

1981) to address those regions in a warm blooded species genome showing a homogenous amount of GC content. Bernardi (1985) described the human genome as a mosaic of different isochore families. Isochore sizes can vary between the size of a gene and the size of a chromosome band (Bernardi 1993). There are five different isochore families known in human genome based on their GC content: L family (L1 and L2) which comprise 62% of the human genome and are the GC poor regions, H1 and H2 which are the GC rich regions and comprise 22% and 9% of the human genome respectively, and H3 which is the GC richest isochore and comprises only 3-5% of the human genome (Saccone 1993; Bernardi 1993). Saccone also showed that low GC isochores correspond to G-bands (Geimsa positive bands) and GC rich isochores correspond to R-bands (reverse bands or Geimsa negative bands).

The biological significance of such isochores is still hotly debated. Bernardi and his colleagues have emphasized the fact that in cold-blooded vertebrates, this banding phenomenon is far less pronounced than in warm-blooded vertebrates. They suggest that the isochore structure is a genomic adaptation to the higher body temperatures among endothermic vertebrates (Bernardi, 1993). The overall range of the GC contents among these five families of isochores is between 30% and 60%.

Because of the very high GC content in H3 Isochores, there is a codon usage bias seen in this region toward those codons with G/C in their third codons positions. This bias can be also observed in amino acid composition, towards amino acids that have G/C in their first two codon positions (i.e. Arginine, Alanine, Glycine, and Proline) rather than amino acids with A/T nucleotides in those codon positions (e.g. Lysine) (Bernardi, Bernardi 1985; Bernardi, Bernardi 1986; Bernardi, 1989). Bernardi also shows (1986) that the ratio of (Alanine + Arginine) / (Lysine + Serine) increases four folds from GC poor to GC rich genes as Alanine and Arginine

are amino acids that most contribute in the thermostability of the proteins. He also shows (1989) that in GC richest isochore (H3) of the human genome, the gene concentration is much higher comparing the other isochore families. I will talk more about the existence and the importance of the isochore structures in the next section where I list some theories explaining the very heterogeneity both between different genomes as well as within a single genome.

1.4. Explaining the Variations in Nucleotide Content between and within Genomes

There are a handful of theories explaining the heterogeneity in nucleotide content both among different species as well as within a single genome. These theories mostly can be divided into two groups; those based on a different mutation rates and those based on the effect of selection on shaping the genomic nucleotide content. There is also another way that these theories can be categorized. Based on the effect of the body temperature on the evolution of the genomic nucleotide content both mutationalist (neutralist) and selectionist theories can be divided into “thermal” (i.e. those relying on the effect of the body temperature) and “non-thermal” (i.e. those not relying on the effect of the body temperature) groups.

Noboru Sueoka (1961, 1962, and 1988) suggests directional mutation pressure as the main cause behind the intraspecies heterogeneity in genomic nucleotide content (specifically at third codon position) in mammalian protein coding genes. He explains that this difference might be basically due to “local structural elements” of the chromatin which can lead to different behaviour of some “mutagenic” molecular machineries such as DNA replication and repair. Sueoka’s theory can be considered as a mutationalist (neutralist) model. This model has been confirmed by some other researchers such as Cox (1972).

On the other side, Bernardi’s “Genomic Stability Model” (1989) which can be considered a “thermal selectionist” model doesn’t rely on directional mutation pressure. In this model

Bernardi and his colleagues explain that in warm-blooded vertebrates, there is a selective advantage of GC-rich isochores because of the higher thermostability of GC-rich DNA. In addition, the GC-rich coding sequences may encode more amino acids that contribute to the thermostability of proteins. Bernardi and his colleagues pointed that the existence of the very GC rich isochore structures in warm-blooded vertebrates (homeotherms) is the result of the selection against low GC regions while in cold-blooded vertebrates (poikilotherms) we cannot see these GC rich isochore structures. However, as we will see later on, Chojnowski and his colleagues (2007) were able to find these GC rich isochore structures in American alligator's genome. This finding was contradictory to Bernardi's genome stability model for the evolution of isochores.

Later in 1999, Lobry and Gautier introduce "amino acid constraints", as a "non-thermal selectionist" model. Based on "amino acid constraints" model, structural or functional constraints or even the localization of the proteins inside the cell can affect the nucleotide content of the genes. For example, the selection in favour of the proteins with higher amounts of hydrophobic amino acids (e.g. Leucine, Methionine, Phenylalanine, Isoleucine, Tryptophan, Valine, and Tyrosine) in *E. coli* in order to be located as integral membrane proteins, can affect the content of the nucleotide of this species' genome in certain genes coding for these proteins. All these amino acids are coded by codons which have either Adenine or Thymine on their second codon position. In some cases such as Leucine, Methionine, Phenylalanine, Isoleucine, and Tyrosine both first and second codon positions are occupied with A/T. However, Lobry and Gautier also mention that these proteins only consist 10% of the total proteins in *E. coli* proteome and thus cannot have a great effect on the total nucleotide content of the genome. Thus, this theory, at least by itself, doesn't seem to be sufficient enough to explain the observed heterogeneities in nucleotide content values.

It is well known that the bias in the codon usage is related to the relative abundance of the specific tRNA (and as a result to the level of the expression). Sueoka (1995) suggested that a selection in favour of those codons with abundant tRNA can be a major player in shaping the nucleotide composition at third codon position in synonymous codons.

However, the bolder and more pronounced variation in the nucleotide composition at third codon positions (which are not affected by selection) suggests that the reason behind this variation might be something beyond only selection; moreover, since these positions are not sensitive to selection, there is a bigger chance of fixation of the mutations at these positions through genetic drift (neutral theory of molecular evolution).

Under the neutrality model (mutationalist model), one might expect the substitution profiles to reflect an equality in the chance of fixation among different types of mutations. But human polymorphism datasets clearly show that this is not, in fact, the case and there are more A::T→G::C mutations comparing the opposite direction (i.e. G::C→A::T) (Eyre-Walker 1999; Smith and Eyre-Walker 2001). This apparently shows a selective advantage towards Guanine and Cytosine alleles over those with Adenine or Thymine. However, Eyre-Walker explains that this inequality is not due to selection and, in fact, is due to variation in mutational pressure.

Eyre-Walker (1999) explains an alternative mechanism shaping the isochore structures called “Biased Gene Conversion towards GC” or BGC(GC). This mechanism, which can be considered a non-thermal mutationalist model, depends on biased DNA repair during recombination. Later, other studies by Galtier and his colleagues (2001) and Duret and his colleagues (2002) also confirmed this model. This model can actually explain the observed correlation between the recombination rate and the GC content in different regions of a chromosome. Gene conversion happens when a minus and a plus strand of two sister chromatids form a heteroduplex during

recombination. If the formed heteroduplex include a mismatch (in a heterozygous genome) this might trigger the DNA repair systems. The mismatch will be corrected and one of the nucleotides in either strand will be replaced by the complementary nucleotide of the other strand. For example if the mismatch is formed between Guanine and Adenine, either Guanine will be replaced by a Thymine to be properly paired with Adenine on the other strand, or the Adenine itself will be replaced by a Cytosine to be properly paired with Guanine. This leads to a change in the proportion of the genotypes of the gametes from a 1:1 proportion to a different proportion. If this repair system is biased towards one of the two possibilities (e.g. towards G::C pairs or A::T pairs), in a germ line with a heterozygous genotype, the number of produced gametes with one of these two alleles will be greater than the other type. For example, if the DNA repair system is biased towards G::C pairs, the number of produced gametes with G::C alleles will be greater than the number of gametes with A::T alleles. This, in turn, will increase the chance of fixation of the G::C alleles in the population. Later in 2001, Galtier showed that in a population with effective size of 10^4 - 10^5 , considering the recombination rate of 10^{-8} and also (up to) several kilobases of a heteroduplex structure, this mechanism can be considered effective enough to significantly influence the fixation process of GC/AT polymorphism and the resulting change in the GC content.

Frank and Lobry (1999) mention “selection for function at nucleotide level” along with “amino acid constraints” and “codon bias” models as three “non-thermal selectionist” models. They explain the functional constraints on the nucleotide level as the abundance of certain types of complementary sites in ribosomal RNA (rRNA) can limit the usage of certain types of codons. They explain the necessity of the rRNA complementary sites as they act as check points to avoid any “slippage” of the mRNA during the translation process and to prohibit any chance of frame

shift mistranslation. For example if the complementary sites in 16S rRNA in a species start with Cytosine, it will be easy to understand the reason of the observed abundance of the codons starting with Guanine. This model which in literature is referred to as SCU (i.e. Selection on Codon Usage) has been studied later on by other researchers in different species. Serres-Giardi and her colleagues (2012) discuss this model in their broad studies on plants. They argue that the SCU model only can explain the variation in GC content at third codon position and partly at first codon position not the overall GC content variation. Although in the same study they show that the variation in GC content at third codon position is much stronger than the other two positions and also they show that the value of the GC content at third codon position is positively correlated with the expression levels. They connect this phenomenon with the abundance of the codons ending with Guanine and Cytosine in plants (Wang and Roossnick, 2006) and conclude that SCU can be considered as one of the causes of the GC content heterogeneity in plants along with other causes they explain in their study.

In contrast to Bernardi's "thermal selectionist" model, Fryxell and Zuckerkandl (2000), proposed a model that can fit as a "thermal mutationalist" model. According to this model which is known as "deamination feedback loop", higher body temperature increases the rate of deamination of Cytosine (that in turn leads to the transition of Cytosine to Thymine), thus leading to a decrease in GC content. They showed that for each 10°C increase in the body temperature, the rate of the Cytosine deamination increases 5.7 folds. Thus, if an ancestral genome is highly GC rich, the deamination cycle can lead to the creation of GC-poor (i.e., AT-rich) regions within the genome, especially in areas that are not subject to the countervailing effect of strong purifying selection (Fryxell and Zuckerkandl, 2000).

Other than the effect of higher body temperature on the rate of Cytosine deamination, they also showed that the change in GC content in a stretch of DNA can also affect the rate of Cytosine deamination. They explain that Cytosine deamination happens 143 times faster in single stranded DNA (i.e. melted DNA). Thus along with activation energy this reaction (i.e. Cytosine deamination) needs local temporary melting of DNA. DNA melting point, in turn, decreases along with the decrease in the GC content of the DNA (because of the difference in the number of hydrogen bonding between A::T and G::C pairs). They explain that while a 10% decrease in GC content can decrease the melting point as much as 4.1°C, a decrease in melting point can have the same effect on the rate of Cytosine deamination as an increase in body temperature. Thus, one can say that a decrease in GC content can lead to an increase in Cytosine deamination, through a decrease in melting point, which in turn causes a decrease in GC content. This shapes a progressive positive feedback loop between the GC content and Cytosine deamination.

They also explain the lack of distinct classes of isochores in fish and amphibians (cold-blooded vertebrates) versus warm-blooded ones using the difference in their average body temperature. They mention that in cold-blooded vertebrates such as amphibians and fish, which their average body temperature is around 20°C, the rate of Cytosine deamination is 20.6 folds slower compared to warm-blooded vertebrates such as birds and mammals with the average body temperature of 37°C. This lower rate of Cytosine deamination leads to an insignificant positive feedback loop between GC content and Cytosine deamination which, in turn, causes the lack of distinct classes of isochores compared to warm-blooded vertebrates. They also mention reptiles as an intermediate species (in terms of body temperature) and explain that they actually show GC rich isochore structures because of higher body temperature compared to amphibians and fish

and conclude that the evolution of GC rich isochores in vertebrates have begun once the early vertebrates started leaving aquatic life style and occupying land.

Xia and his colleagues (2002) studied the effect of increasing temperature on the frequency of nucleotides and dinucleotides in *Pasteurella multocida*'s genome. In their study, they cultured this bacterium for over 14000 generations under increasing temperatures to attenuate the highly virulent bacterium to its vaccine form. Interestingly, they found that, despite the popular hypothesis that correlated high temperatures with high GC contents, in their bacterial cultures, the content of Guanine and Cytosine showed a decrease with increasing temperature. They proposed two hypotheses explaining this observation; first they explained this using the Cytosine deamination model in which increases with higher temperatures. They backed their hypothesis with the fact that the content of AT at third codon position is higher than the first two codon positions, which in turn, shows the effect of mutational pressure in making more ATs, since the third codon position is less selectively constrained than the first two codon positions. Secondly, they explained this decrease in GC content based on selectionist model. They explain that based on the recent studies, that show the conformation of DNA is heavily dependent to the dinucleotide, trinucleotide, and tetranucleotide elements, those elements that are rich in A and T help the stability of the DNA conformation. Thus it will not be a surprise to observe an increase in AT content in higher temperatures (Xia *et al.* 2002).

In their successful search for GC rich isochore structures in American alligator's genome, Chojnowski and his colleagues (2007) challenge Bernardi's "genomic stability model". In their report, they also unify different existing models of isochore evolution into a framework based on the main evolutionary forces behind the evolution of isochores. They divide these models into two main categories labelled as mutationalist (or as it is referred to by others, neutralist) and

selectionist models. As explained earlier in this section, the former is based on the variable mutation rates across genomes (as well as between genomes) and is consistent with the neutral theory of the molecular evolution and the latter states the role of natural selection on shaping the isochore structures (majorly through selection against low GC regions in homeotherm species).

Although the very existence of the isochores in the human genome and the causes behind the evolution of these structures are still matters of controversy (Nekrutenko and Li, 2000; Bernardi 2001; Häring and Kypr 2001; Cohen et al. 2005; Constantini et al. 2006), but even if isochores, as originally defined by Bernardi, do not actually exist, there is a general agreement that some form of heterogeneity in nucleotide content clearly exists within genomes of vertebrates.

1.5. Phylogenetic inference is affected by nucleotide bias

Other than codon usage bias and amino acid content, the bias in nucleotide content can also affect phylogenetic inference studies. Different methods of phylogenetic inference use different models of DNA/Protein molecular evolution. These models for molecular evolution of DNA basically define the rate of the substitution between different nucleotides between two sequences (some of them are even site specific) as well as the frequency of each nucleotide. Thus the frequency of nucleotides can directly affect the inference of phylogenetic relationship between different species (in DNA based methods) or through affecting the frequency of different amino acids (in protein based methods).

In Section 1.2 (i.e. GC content variation between genomes), I briefly explained the effect of variation in genome nucleotide content on the phylogenetic studies when I explained Foster and Hickey's study (1999) on the effect of the nucleotide content variation on phylogenetic inference methods. Foster and Hickey examined both nucleotide sequence and amino acid sequence based methods of inferring phylogenies. They showed that bias in nucleotide content not only can

affect phylogenetic inference through DNA based methods, but despite what that was shown previously by Hesagawa and Hashimoto (1993) about the reliability of protein based methods when there is a strong bias in nucleotide content, this bias can, in fact, affect the phylogenetic inference through protein based methods as well.

Researchers such as Hesagawa and Hashimoto, who believed that in the case of bias in nucleotide content, a protein based approach in phylogenetic inference can be more reliable, base their claims on the greater functional constraints in keeping protein compositions constant. However Foster *et al.* previously had shown (1997) that along with codon usage, the amino acid content, as well, can suffer from a bias in nucleotide content. This effect was, as well, shown before by some other groups (Muto and Osawa 1987).

Foster and Hickey (1999) examined mitochondrial genes of eight different animal species covering a broad phylogenetic range and different nucleotide and amino acid composition with a known consensus phylogenetic relationship. Their list of species included Honeybee, Nematode, Locust, Fruit fly, Brine Shrimp, Allomyces, Sea Urchin, and Chicken. In their list, two species; honeybee and nematode shared the properties of high bias in their nucleotide and amino acid content. Especially the ratio of those amino acids coded by AT rich codons (i.e. Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine, and Lysine) to those amino acids with GC rich codons (Glycine, Alanine, Arginine, and Proline) was the highest among all the other species. Foster and Hickey reconstructed the phylogenetic relationship between these species using Maximum Likelihood method using DNA sequences. In their phylogenetic inference results, honeybee and nematode were grouped together regardless to their lack of biological relatedness. In order to avoid the bias arising from codon bias they also used different methods of phylogenetic inference using protein sequences, such as Maximum Likelihood, LogDet distance

correlation, distance methods, and parsimony methods. These methods although failed to infer the correct phylogenetic tree by grouping honeybee and nematode. Interestingly, the bootstrap values in all cases were showing high confidence in the resulting trees. Even with the removal of the AT and GC rich amino acids (i.e. Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine, Lysine, Glycine, Alanine, Arginine, and Proline) from the alignments, they could not fix the problem of grouping honeybee and nematode. The only difference was in lowering the bootstrap values. They explained this as follow; an ancestor with a neutral AT/GC amino acid content did not gain more AT rich amino acids from converting all the other GC rich amino acids, and on the other side, did not lose all the GC rich amino acids converting all of them to AT rich amino acids. They explain that these gaining AT or losing GC content has also affected the “AT/GC neutral amino acid content as well”. In other words they explain that the effect of the bias in nucleotide content is not exclusively limited to AT/GC rich amino acids but it can be observed in “AT/GC neutral” amino acids as well. This is why even removing these “troublesome” amino acids is not enough to infer a correct phylogenetic tree.

There are a handful of other studies mentioning how bias in nucleotide content can interfere with the correct inference of phylogenetic studies. In a more recent study, Benoit Nabholz *et al.* (2011) studied the phylogenetic relationship among birds (Neoaves) using 1995 genes, (brain transcriptome), benefitting from the next generation sequencing technologies. In their studies, Nabholz and her colleagues test the effect of the observed heterogeneity in the GC content between different lineages of their model organisms (they found the excess of Guanine and Cytosine content at third codon positions of some of the species with the strongest effect in passerines). To find out about the effect of the GC content on the phylogenetic inference, Nabholz and her colleagues split their datasets to two datasets; one only including the third

codon position and the other including the other two codon positions (i.e. first and second codon positions). The phylogenetic inference results, from a Maximum Likelihood method, on the two datasets showed different positions for budgerigar and also the relationship between zebra finch, blue tit, and pied flycatcher. In their study they showed that despite to the large number of genes they used for reconstructing the phylogenetic tree, the inference of the phylogenetic relationship between different species can be still affected by the nucleotide content of different species in the dataset. They explain that third codon position specially has a great effect. In their study, when they included the total sequence, the resulting tree did not position budgerigar next to passerines as it did when they removed the third codon position.

One of the most complicated phylogenetic puzzles for researchers in recent years is the phylogenetic relationship within the malaria parasite lineage. Today, our knowledge about different biological aspects of these parasites such as their genome sequence (Gardner *et al.* 2002, Carlton *et al.* 2008), their proteome, and their life cycle is very vast; however, the evolutionary relationship between different species within this genus is yet to be clearly defined.

There are a handful of theories about the origin of each of these species especially those involved in human malaria disease. Elizabeth Pennisi (2001) wrote a short note in Science magazine where she mentioned an ongoing debate between two groups of scientists about the origin of malaria parasite *P. falciparum*. Pennisi puts them into two groups; one group are those who believe that *P. falciparum* was evolved around the time the hunter human started settling down and developing agriculture about several thousands years ago, and the second group believe that this species has been diverged from its sister species *Plasmodium reichenowi* around the same time their host species (human and chimpanzee respectively) have done around 8 million years ago.

In another study, Dávalos and Perkins (2008) used genome scale dataset (>100 genes) to infer the phylogenies of eight *Plasmodium* species. They showed that phylogenetic studies within genus *Plasmodium* are strongly affected by nonstationary in base frequencies, especially at third codon position. In a sequence, nucleotide substitution model is considered stationary when the nucleotide composition of that sequence stays constant over time (Gu and Li 1998), thus if the composition changes over time, it will be considered as a nonstationary substitution model. They claim that codon-partitioning methods, which consider different rates of changes in different codon positions, as well as protein based methods, which exclude the biased base frequencies, can recover the phylogenetic studies more accurately. But Foster and Hickey (1999) had shown before that even protein based phylogenetic studies can suffer from the bias in nucleotide content through changes in the content of amino acids.

Hayakawa *et al.* (2008) explained that host-switch events might have been a strong trigger driving the evolution of different species within genus *Plasmodium* as an adaptive trait of the parasites. They also suggested a bird or reptile malaria parasite as the most likely common ancestor for the malaria parasites (at the root of the phylogenetic tree) including those infecting mammalian species. But later on, Silva and her colleagues (2010) confirm the association between malaria parasites and their hosts. Analyzing 45 genes, they showed that two *Plasmodium* species, *P. falciparum* and *P. reichenowi* have diverged from each other around the time that human and chimpanzee diverged from one another. They also showed that *Plasmodium vivax* and *Plasmodium knowlesi* have diverged from each other around the time their two host species (Great Apes and Old World Monkeys respectively) have separated their evolutionary paths.

In a different study however, Blanquart and Gascuel (2011), using mitochondrial sequences, claimed that there is a common origin for both rodent malaria parasite and those of the great ape lineage (including *P. falciparum*).

As it is mentioned above, different groups have studied the phylogenetics of *Plasmodium* genus using one or a few nuclear or mitochondrial genes as the bearers of phylogenetic signals. Genes such as ribosomal RNA genes (Escalante and Ayala 1994), Cytochrome b (Perkins and Schall 2002), a small set of mitochondrial genes (Blanquart and Gascuel 2011), or even the total mitochondrial genome sequences (Hayakawa *et al.* 2008) have been analyzed in these studies. But yet researchers have not reached a consensus phylogenetic tree for this lineage. It will not be a surprise if one thinks that this complication in the phylogenetic inference within this lineage and the strange behaviour of their genomic nucleotide content might be somehow related, as Dávalos and Perkins (2008) pointed it out. At least, one might help explaining the other one.

Other than the strange behaviour of the genomic nucleotide content in this lineage, as we will see later, in this lineage, especially between the two species responsible for human malaria (i.e. *P. falciparum* and *P. vivax*), I will study the rate of the substitution between different nucleotides to see if it follows the usual trend. As explained earlier in this chapter, the rate of substitutions between different nucleotides is one of the attributes defined by different models of the molecular evolution of DNA. Thus, other than the strange behaviour of the genomic nucleotide content in this lineage, phylogenetic inference can be jeopardized by the unusual rate of the substitution between different nucleotides. To explain this “unusual” substitution rates, I first will explain the “usual” substitution rates both between and within different classes of nucleotides.

I mentioned earlier in Section 1.1, nucleotides can be divided into two classes based on their chemical structure; Purines and Pyrimidines. Changes between two nucleotides within the same class are called transition (*ts*) and changes between any two nucleotides from different classes are called transversion (*tv*) (Wakeley 1996). This means that there can be four different types of transitional changes as opposed to eight different types of transversions.

Although the number of possible transversions is twice as great as the number of possible transitions, the frequency of transitional substitutions is shown to be greater than the other type of changes, i.e. transversion (Figure 1.1) (Vogel, F. and Rörhborn, G. 1966, Fitch, W.M. 1967, and Wakeley J. 1996).

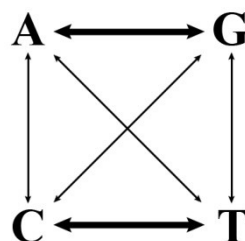


Figure 1.1. Schematic representation of different nucleotide substitutions between four different nucleotides. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

The abundance of transitional changes over transversions (transition bias) can be explained according to the chemical structure of the nucleotides, i.e. a purine can replace another purine simply because of Watson and Crick double helix structure of the DNA (a purine always pairs with a pyrimidine and vice versa) (Watson and Crick 1953; Wakeley 1996). Also, the very high frequency of deamination of Cytosines to Thymine in CpG islands (a pyrimidine replacing another pyrimidine), especially in single stranded DNA can be considered as another explanation

behind the higher frequency of transitions versus transversions (Frank and Lobry 1999). This happens with a higher rate (140 times faster) during replication when the lagging strand is left single stranded for a short while (Frederico *et al.* 1990). Keller and her colleagues (2007) studied the methylation of Cytosines in CpG islands in pseudogenes of grasshopper and showed that if we account for these transition mutations, there will be no more significant transition bias left. Echols and Goodman (1991) showed that the G/C \rightarrow A/T transitions (G::T, A::C mispairs) have the highest frequency among mutations in *E. coli*. Thus one can conclude that the bias towards transitions basically starts to happen at the level of mutation.

Crozier and Crozier (1993) also showed that the bias towards transitions at third codon position in highly diverged coding sequences can be due to the conservation of the coded amino acid sequences. Another explanation is based on Kimura's neutral theory of molecular evolution and the rate of mutation fixation through genetic drift. This idea argues that the fixation rate in the codons subject to transitional changes at third codon position differs from those codons affected by transversions, and this can be one of the causes behind the observed transitional bias (Wakeley 1996). Looking at the codon table, one can simply calculate that only 3% of the transitional changes at third codon position can affect the type of the encoded amino acid. But this amount for transversions is as high as 41% (note that I exclude stop codons in my calculations). The higher rate of changes in the coded amino acid increases the possibility of elimination from the population through selection and as a result it decreases the chance of fixation by genetic drift. In other words, there is a bigger chance for transitional changes at third codon positions to be fixed in the population than transversions. These amounts (i.e. the percentage of changes leading to change in the encoded amino acid) for mammalian mitochondrial DNA are as low as 0% for transitional changes and as high as 43% for

transversions (Wakeley 1996). Thus one might say that while almost half of the transversional changes at third codon position in mammalian mitochondrial genomes lead to changes in the coded amino acids (and as a result increase the susceptibility to selection), none of the transitional changes at this codon position changes the coded amino acid and thus these types of substitutions get the chance to be fixed through genetic drift. These numbers are also correct when we take the invertebrate mitochondrial genomes and their corresponding codon table into account (i.e. all of the transitions at third codon position are synonymous mutations while over 43% of the transversions at third codon position are replacement mutations). The standard codon table as well as the mammalian mitochondrial and also the invertebrate mitochondrial codon tables are provided as Supplementary Figure S.1. It is clear that if a mutation leads to a change in the coded amino acid, the resulted protein can be either harmful or neutral, or even in rare cases beneficial to the fitness of the organism. Considering the very lower rate of beneficial mutations comparing to the harmful or neutral ones (Sniegowski *et al.* 2000), the higher rate of replacement mutations (mutations leading to change in the encoded amino acid) in an organism can eventually increase the chance of the elimination of the organism from the population through selection against harmful mutations. Thus one can conclude that evolution also favours the accumulation of transitional changes over transversions since there is less possibility of change in the type of the coded amino acids when the changes at third codon positions are transitions. In other words transitions at third codon positions in mammalian and invertebrates mitochondrial genomes do not change the fitness of the organism, thus have been the favoured type of changes through the course of the evolution.

In conclusion, one can say that the bias towards higher transition rates has been happening in different levels throughout evolution (i.e. mutation bias and selection against replacement

mutations). In literature, the aforementioned bias towards transitions is usually represented as the ratio of the rate of transition over the rate of transversion (i.e. ts/tv).

Brown and his colleagues (1982) studied a set of five primate partial mitochondrial DNA sequences and showed that the rate of transitions in these mitochondrial genomes is much higher than the rate of transversions. They also showed that when we compare two very closely related species (e.g. human and chimpanzee), 92% of the changes are transitional type. This value falls to as low as 45% when they compared two farther species (e.g. primate and non-primate species) as a result of multiple changes in one nucleotide site (because of the longer divergence time they have had enough time for multiple changes to occur at the same site). They also showed that not only the rate of evolution in mitochondrial genomes is much higher than nuclear genomes also the bias towards transitions is much higher in these genomes comparing the nuclear genomes. These studies clearly show that to have a better understanding about the complexity of the phylogenetic relationships within *Plasmodium* lineage, as well as other unsolved phylogenetic inference puzzles, one should study the evolution of the nucleotide content of the species within that specific lineage very closely. This, in fact, was one of the main motivations driving my research. I will discuss my research goals and hypotheses in the next section.

1.6. Thesis hypothesis and goals

In this research I mainly focus on the heterogeneity both between genomes as well as within a single genome. In this study I try to pinpoint how this heterogeneity can affect different aspects of a living organism as well as how it can affect different evolutionary studies such as phylogenetic studies. In the next two chapters of this thesis I will focus on the variation in the genomic nucleotide content between different genomes. In Chapter four, however, I look deeper

into the genomic nucleotide contents when I investigate the variation in nucleotide content within a single genome.

In Chapter two I study the variation in the genomic nucleotide content in the nuclear genomes within the genus *Plasmodium*. Different species within this genus show a significant bias in their nucleotide contents towards Adenine and Thymine. Although among these species, there are a few that apparently have escaped this bias that runs through the entire genus. In this chapter I investigate this “escaping” from bias towards AT in one of these species that shows a fairly unbiased amount of AT content (i.e. *P. vivax*). My goal in this chapter is to reveal if these species have actually escaped the AT bias or the story has a different twist to it.

To be more specific, in Chapter two, I study the variation in genomic nucleotide content between this species and another species with an AT biased nucleotide content. For this purpose, I choose *P. falciparum* species which is also responsible for human malaria (the same as *P. vivax*). I compare the changes in their nucleotide content since their divergence, around ten million years ago, to find out what has been happening in these genomes in terms of changes in their nucleotide contents. For this purpose, I follow Frank and Lobry’s first methodology (mentioned earlier in this chapter). I calculate the nucleotide content of the two *Plasmodium* species’ genomes. I also compare the nucleotide content in the orthologous subset of coding sequences between these genomes by aligning the two genomes against each other. In order to study the direction and the pace of the changes in the nucleotide content in these two species since their divergence, I further analyze the nucleotide content in conserved and variable sites both separately and combined. I repeat this analysis in all three different codon positions. The conserved sites between these two species will be taken into account as signals from their most recent common ancestor to explain the direction of the changes since the divergence of the two

species. I also investigate the variable sites between these two genomes to be able to follow the direction and pace of the changes in the nucleotide content in each of the two genomes. This might give us an idea on why phylogenetic studies within this genus has been a puzzle for scientists for years.

In Chapter 3, however, I use the second methodology mentioned by Frank and Lobry to study the nucleotide content variation by building the substitution matrices between orthologous genes of different genomes. I investigate the nucleotide substitution pattern in mitochondrial genomes of the same two species. In this chapter I investigate the presence and the direction of the bias in nucleotide content in the mitochondrial genomes of these two species. Moreover I will examine the effect of the probable bias and its direction in these two genomes on the rate and direction of changes between different nucleotides between these two mitochondrial genomes. My final goal is to see how bias and its direction in each of these two genomes might affect the rate and direction of substitutions between different nucleotides. Moreover, I investigate the possibility of other causes behind substitution trends I see between different nucleotides of the two genomes. For this purpose, I investigate the possibility of saturation effect on the different substitution rates by comparing one of my model organisms (*P. falciparum*) with a very closely related species (i.e. *P. reichenowi*).. My final goal in this chapter is to show the possible effect of the nucleotide bias in two genomes on the pattern of substitutions between different nucleotides both when the direction of the bias in the two genomes is the same and when they show opposite directions. I investigate these patterns in both transition/transversion levels as well as within each of these substitution types.

In Chapter four, I look deeper into the genomes when I study their nucleotide content by comparing different parts within a single genome. For this purpose, I choose a genome with two

subsets of genes with clearly different nucleotide contents (i.e. honeybee). Previously, it was shown that, in a genome with two subsets of genes with different nucleotide contents (rice genome), there is a significant difference between the average lengths of the two subsets of genes. By comparing the average length of the genes in the two subsets of high and low GC genes in honeybee's genome, I try to reveal one of the possible mechanisms behind the observed biases in the nucleotide content in these two subsets of genes. I further investigate the possible effects of the variation in GC content on amino acid content of the coded proteins.

Chapter 2. Rapid reversal of nucleotide bias within the genus *Plasmodium*

2.1. Introduction

In Chapter 1, I discussed that although there is a general tendency for closely related species to show similar values of their genome nucleotide content, in fact, this is not always the case. I mentioned two *Plasmodium* species; *P. falciparum* with unusually low GC content of only 19.44% and *P. vivax* with fairly normal value of GC content (44%) as examples for the latter statement. In this chapter we will have a closer look at what has been happening in terms of the evolution of genomic nucleotide content in these two species since their divergence from a common ancestor. Looking at the values of the GC content between these two species, the first and naive conclusion one might come up with, could be that it is *P. falciparum* which has undergone a dramatic change in its genomic content since its divergence from *P. vivax* millions of years ago. In this chapter I seek for more accurate signals in the genomes of these two species in order to verify if this is, in fact, the story.

Our findings might be helpful for the ongoing debate on the origin of the *P. falciparum* as Elizabeth Pennisi (2001) mentioned in her short note in Science magazine where she divided the current theories into two groups, those who state that *P. falciparum* was evolved when human started developing agriculture several thousands of years ago, and the those that state this species has been diverged around 8 million years ago from *Plasmodium reichenowi* when their host species diverged from each other. As I mentioned in Chapter 1 the bias in nucleotide content of the genes used for reconstructing the phylogenetic relationship within the genus *Plasmodium* affects these studies since different species within this genus cover a wide range of GC content values. Thus one might suggest that having a broader yet more accurate understanding of the

mentioned biases and their directions can be helpful correcting our current phylogenetic inference methods and getting more realistic results revealing the history of this genus.

Thus basically in this chapter I mainly focus on the comparison of the nucleotide content of each of these species genomes. I go further and calculate these values in different codon positions in different sites (e.g. variable sites, conserved sites) in order to find any signal that can explain the direction of the changes happening in the nucleotide content in these species through the course of evolution. The reason I calculate these values in different codon positions as well as different sites is the fact that the pace of accumulation of the changes is different between different codon positions as well as between different sites. This is due to the selection pressure which is different in different sites as well as different codon positions. Synonymous codon usage at third codon position has made this codon position the least sensitive to the selection while the two first codon positions (specially the second codon position) are more sensitive to any changes. Also since variable sites are those sites that changes actually have happened, calculating the changes in these sites should give me more pronounced results comparing to when I include the conserved sites in my calculations. Conversely, the conserved sites provide an indication of the nucleotide composition of the common ancestral sequence. These different paces of accumulations of changes in different sites and codon positions might give us a clue about the direction of the changes. This can be explained simply if we assume that conserved sites that have not changed since the divergence time of the two species and thus can be considered as a common signal from the most recent common ancestor of the two species. Thus if we compare the nucleotide content of the variable sites in two species with their so called “ancestral signal”, which was concluded from the conserved sites between the two species, we

can find a clue about the direction and also the rate of changes between different nucleotides between these two species since their divergence.

This explains the big picture of my methodology in this chapter. Surprisingly the results of my study reveal a counterintuitive, odd, and unexpected behaviour of the genomes of these two species in terms of the direction of the bias in their content. More specifically about *P. vivax* we will see that the content of this genome shows a back and forth (I use the analogy of ebb and flow in the title of the thesis) behaviour through the course of its evolution. Briefly put, we will see that the genome content of the most recent common ancestor of *P. falciparum* and *P. vivax* shows a bias towards a very low GC content meaning that the content of its genome has been losing its GC content before the two species diverge. But after the divergence of these two species, *P. vivax* starts gaining back its GC content at a very fast pace.

2.2. Methods

In order to study the evolution of the nucleotide content within the genus *Plasmodium*, I identified six species for which the entire genomes have been sequenced and fully annotated: *P. vivax*, *P. falciparum*, *P. knowlesi*, *P. berghei*, *P. chabaudi*, and *P. yoelii*. I then calculated the nucleotide content of the coding sequences for each species (Table 2.1).

Among all the six species in Table 2.1, the two species *P. falciparum* and *P. vivax* which have the two extreme nucleotide contents and infect human were the main focus of my interest. I calculated the distribution of the GC content in coding sequences in these two species. To calculate this distribution, I downloaded all the coding sequences from both *P. falciparum* and *P. vivax* and trimmed the datasets for very short (shorter than 150 bp) and very long sequences (longer than 30,000 bp) and also for those sequences not starting with the start codon ATG.

The distributions in Figure 2.1(A) are calculated based on the nucleotide count of 5491 *P. falciparum* coding sequences and 5435 coding sequences from *P. vivax*. To produce these histograms, I used a bin size of one percent (1%) of GC and counted the number of coding sequences falling in each category (bin). For example, the data point (23, 604) on *P. falciparum* histogram in Figure 2.1(A) shows that there are 604 coding sequences in which their GC content is equal or greater than 23% and smaller than 24% and so on. Also, to investigate further about the observed opposite direction of the two distributions and to make sure if this difference is due to either a species specific set of genes in either organism or it can be seen through the whole genome, I extracted the orthologous subset of genes between these two species and calculated the distribution of GC content in these two orthologous subsets of genes.

To create the orthologous datasets I aligned the 5491 *P. falciparum* protein sequences against 5435 protein sequences from *P. vivax* using standalone BLAST (Altschul *et al.* 1990). I used

protein sequences for the alignment to avoid any bias arising from codon degeneracy affecting the homology search. Those hits with expectation values larger than $1E-20$ and low alignment scores have been left out from the “all against all” alignment results dataset. Then I assigned the corresponding annotated coding sequences from both *Plasmodium* species to their protein sequences. The resulting dataset contains 4283 orthologous pairs of genes between *P. falciparum* and *P. vivax*. The distribution of the GC content of the remaining datasets were calculated the same way as I calculated the distribution of the non-aligned datasets of coding sequences. Figure 2.1 (B) represents the distribution of orthologous set of coding sequences between *P. falciparum* and *P. vivax*.

I also further examined the orthologous dataset in order to see if the differences between GC content values in this dataset is a characteristic of the entire genome or is it due to the effect of a subset of unusually long genes with extremely low or high GC contents. In order to do this, I performed pairwise alignment, using stand alone ClustalW (Larkin MA. *et al.* 2007), on each pair of the proteins in the orthologous protein dataset (4283 orthologous pairs) and then removed the gap columns in order to have an even more fair set of data in terms of having the same number of amino acids. Then I assigned the nucleotide sequences to their corresponding protein sequences and removed the codons corresponding to the previously removed amino acids (gap columns). Thus in the resulting set of orthologous pairs, in each pair, both sequences had the same length and same total number of nucleotides since they have been trimmed for flanking loose ends as well as for gaps. I further used this same dataset for all the analyses on conserved and variable sites, as well as different codon positions.

Next, I plotted all the 4283 data points in a way that on one axis I used the GC content of one of the coding sequences from each orthologous pair and plot it against the GC content of the

other sequence from the same orthologous pair (Figure 2.2). Thus each data point in Figure 2.2 corresponds the GC contents of a pair of orthologous coding sequences between the two *Plasmodium* species. I plotted GC contents of the *P. falciparum* on the X-axis and *P. vivax* on the Y-axis. The dashed diagonal line in the plot represents those points with equal GC contents in both *P. falciparum* and *P. vivax*. If a data point happens to appear above the dashed diagonal line, it means that for that pair of orthologous coding sequences, the GC content of the sequence from *P. vivax* is higher than the GC content of its orthologous sequence from *P. falciparum*. This is the opposite about the area below the diagonal line. Also, to have a closer look at the distribution histogram (Figure 2.1.B), I plotted the GC content of each coding sequence in every single pair of orthologous genes against their alignment length as a common attribute between the two sequences of an orthologous pair. The results are shown in Figure 2.3 (A).

I further examined sequence pairs in the overlapping part of the two distribution clouds in Figure 2.3 (A) to see if the *P. falciparum* sequences orthologous to the subset of *P. vivax* sequences with very low GC content still show lower values of GC contents comparing their corresponding sequences in *P. vivax*. In order to do this, I ranked the whole dataset of orthologous pairs based on GC content of the *P. vivax* coding sequences. Then I picked all the sequences with values less than 30% GC from *P. vivax* sequences (14 sequences) and plotted them along side with their corresponding orthologous sequences from *P. falciparum* against their alignment length (Figure 2.3.B). Again I chose the alignment length as the X axis because it is the same for both of the sequences in each pair of orthologous coding sequence.

To create all these graphs I analyzed every single pair of orthologous sequences in both *P. falciparum* and *P. vivax* (4283 pairs of orthologous sequences). Tables 2.2, 2.3, and 2.4 show the datasets I created and used for my calculations for the GC content in more details (tables here

show only a part of the complete dataset. The complete datasets are provided as supplementary materials in the form of separate spread sheet files). The GC content amounts in these tables are calculated after the sequences were edited for gap columns. The values for GC contents are rounded to one percent.

To understand the sequence of events in either *Plasmodium* species' genomes, I calculated the GC content values in both conserved and variable sites (in all three codon positions) in the two genomes. The results for this calculation are represented in Figures 2.4 (A, B, and C).

Also to investigate further about which nucleotide has had the biggest effect in the observed trend in Figure 2.5, I calculated the content of each of the nucleotides separately. I did this analysis in all three codon positions in conserved sites, variable sites, as well as total sites (conserved and variable sites combined). The results are shown in Figures 2.5 (A, B, and C). I used the same data as in Tables 2.2 to 2.4 to create these figures.

2.3. Results

P. falciparum holds the record for the species with the lowest GC content among all the species with known genome sequence, while *P. vivax* shows a fairly unbiased amount of GC content and the highest among the six species listed in the Table 2.1. Figure 2.1.A shows the distribution of the GC content in coding sequences in these two species.

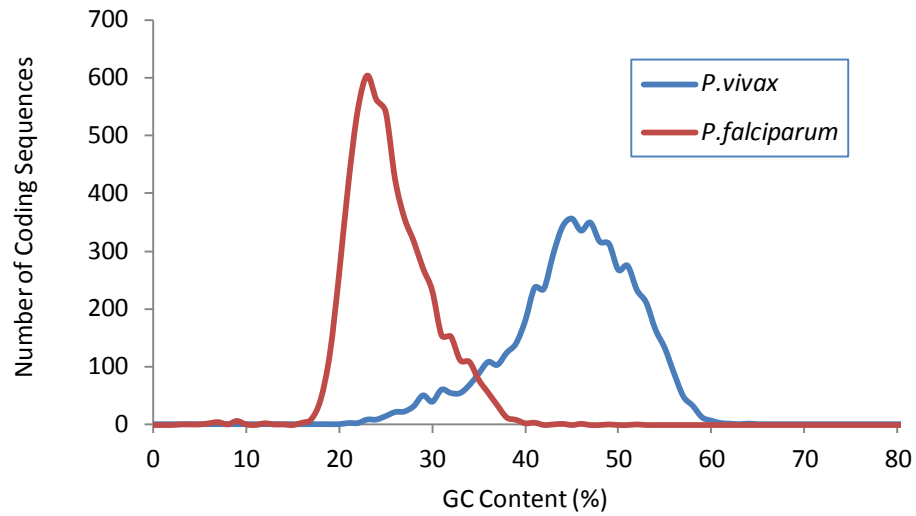
Table 2.1: GC content of six *Plasmodium* species

Species	GC Content (%)	Host
<i>P. vivax</i>	46.2	Human
<i>P. falciparum</i>	23.8	Human
<i>P. knowlesi</i>	40.2	Long-tailed Macaques
<i>P. berghei</i>	23.7	Murine rodents from Central Africa
<i>P. chabaudi</i>	25.5	Murine rodents from Central Africa
<i>P. yoelii</i>	24.2	Murine rodents from Central Africa

These values are based on the GC content of the coding sequences (i.e. introns and intergenic non-coding sequences are not included).

As it is shown in Figure 2.1 (A), *P. falciparum* shows a distinguishable skew towards low GC contents with a peak between 22% and 23% while *P. vivax* shows a fairly bell shaped (although very slightly skewed towards higher GC amounts) unbiased normal distribution around 45%. It is clear in Figure 2.1 (B) that even in the orthologous set of sequences we are still able to see that the GC content distribution in *P. falciparum* is skewed towards lower amounts of GC while in *P. vivax* this distribution shows a fairly normal bell shape around the unbiased amount of GC content (i.e. 45%). This shows that the difference seen between GC content distributions between the two species is not due to a species specific set of genes and the observed variation in GC content runs through the whole genome in almost the same fashion.

A)



B)

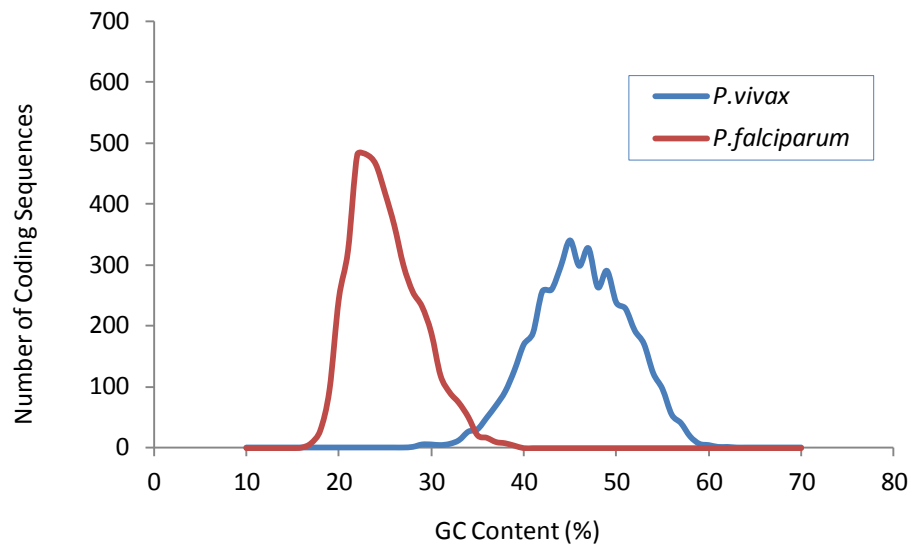


Figure 2.1: Distribution of GC content in coding sequences of two *Plasmodium* species. **(A)** Data for all genes **(B)** Data for the orthologous subset of coding sequences only.

As mentioned in the methods part, in order to see if this trend runs through the entire genome in both species and is not an artefact due to a specific set of very short or long genes with extremely low or high GC content, I plotted each pair of orthologous genes against each other

based on their GC contents. As it is clearly shown in Figure 2.2, all the data points are located above the dashed diagonal line. This means in all the pairs of orthologous sequences, the GC content value in *P. vivax* sequence is higher than its corresponding orthologous sequence in *P. falciparum*. In this figure, we can also see that the resulting cloud of all the 4283 data points is slightly skewed from the bottom left corner to the top right corner. This can be interpreted as those sequences with low GC content in *P. vivax* have orthologous sequences with relatively lower GC contents comparing the rest of the dataset in *P. falciparum* and vice versa.

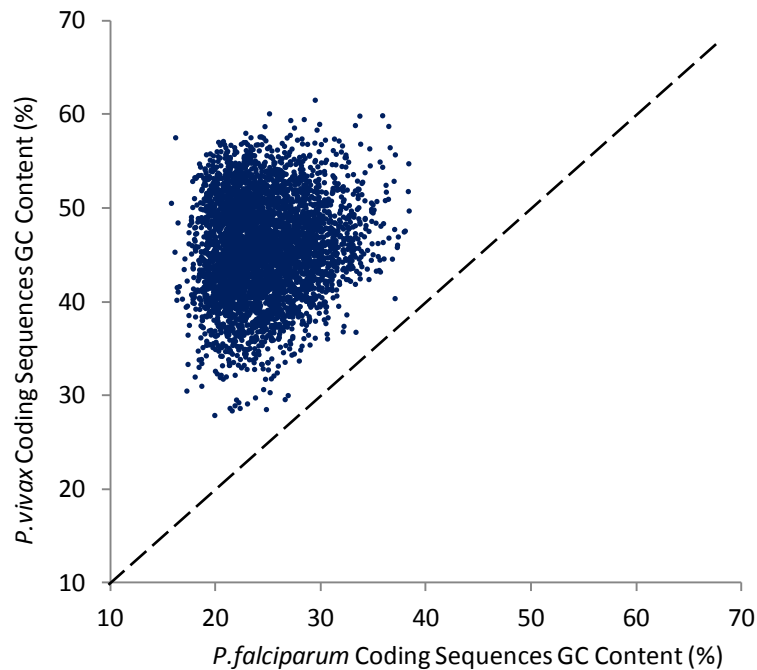
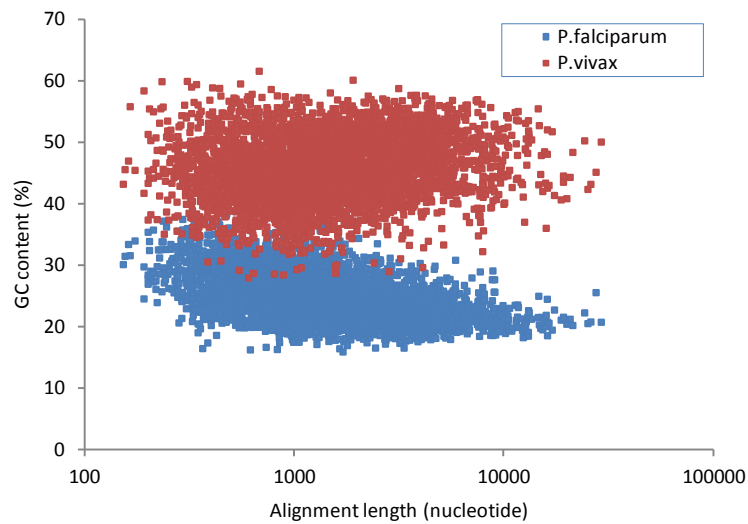


Figure 2.2: The distribution plot of the GC content of 4283 pairs of orthologous genes between the two *Plasmodium* species.

To investigate more about the distribution histogram in Figure 2.1 (B), I plotted the GC content of each pair of orthologous coding sequences against their alignment length (Figure 2.3.A). It is clear that the two clouds of genomes are separate from each other. The *P. vivax* cloud is clearly above the *P. falciparum* one but a small part that these two clouds overlap.

A)



B)

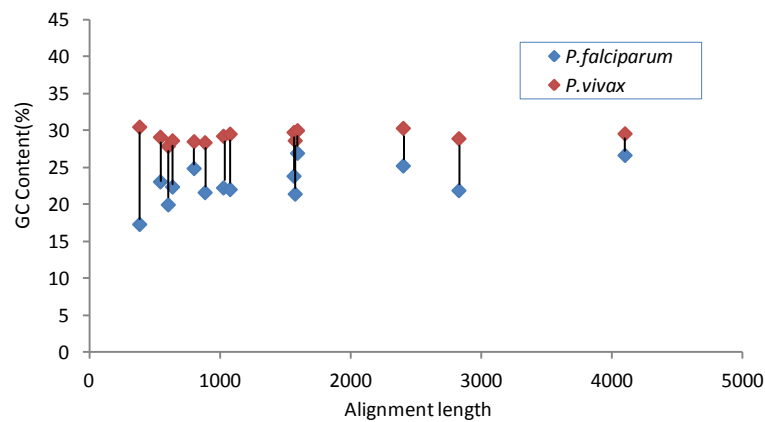


Figure 2.3 GC content distribution of the orthologous pairs in the two *Plasmodium* species plotted against their alignment length (A) Data for all the orthologous set of genes (B) Data for the 14 lowest GC content *P. vivax* (GC content less than 30%) and their orthologous genes in *P. falciparum*.

I used the same dataset used for Figures 2.1 (B) to produce Figure 2.3 (A). In order to have a more comprehensive figure, I displayed the alignment length in logarithmic scale. The overlapping part of the distributions in Figure 2.3 (A) covers the very low GC content subset of sequences from *P. vivax*. As I discussed about the slightly skewed cloud in Figure 2.2, the low GC content subset of genes in *P. vivax* have homologous sequences in *P. falciparum* with relatively lower GC contents comparing the rest of the *P. falciparum* dataset.

As mentioned earlier, in order to see if those sequences from *P. falciparum* that are orthologous to the very low GC content subset of *P. vivax* sequences still show lower GC contents comparing their corresponding sequences in *P. vivax*, I examined sequence pairs in the overlapping part of the two distribution clouds shown in Figure 2.3 (A). The results of this comparison are shown in Figure 2.3 (B). Each vertical line in Figure 2.3(B) connects two orthologous coding sequences from the two species. This figure shows that in all the cases, the sequences in *P. falciparum* still show a lower GC content than their corresponding orthologous sequences in *P. vivax* even though the *P. vivax* sequence has an extremely low GC content. As I mentioned earlier, the datasets I used in my analysis contain 4283 pairs of orthologous sequences. Tables 2.2 to 2.4 show a part of these datasets I used to produce the figures and tables in this chapter.

Table 2.2: A part of the table containing the data for GC content of the total set of orthologous coding sequences (4283 coding sequences) in two *Plasmodium* species

Gene Pair	Gene ID		GC Content (%)	
	P.fal	P.viv	P.fal	P.viv
1	MAL7P1.91	PVX_000525	23	42
2	MAL7P1.89	PVX_000530	20	45
3	MAL7P1.88	PVX_000535	24	42
4	PF07_0074	PVX_000540	24	51
5	PF07_0073	PVX_000545	26	52
...
...
...
4281	PF11_0096	PVX_252300	28	44
4282	PFB0480w	PVX_253300	20	35
4283	PFL1145w	PVX_254300	27	51

The data represented here are based on only coding sequences. The entire table with complete number of rows is provided as supplementary material.

Table 2.3: The GC content and number of conserved sites between the two *Plasmodium* species

Gene Pair	Alignment length	Number of Conserved sites	GC Content (%)
1	7887	3640 (46%)	21
2	15009	9193 (61%)	23
3	1332	860 (65%)	26
4	3924	1551 (40%)	25
5	1590	987 (62%)	34
...
...
...
4281	1005	776 (77%)	33
4282	942	617 (65%)	20
4283	918	578 (63%)	35

This table represents only a part of the complete table containing the data of the alignment of the two entire genomes (the same dataset I used for Table 2.2). The data represented here are based on only conserved sites in orthologous coding sequences. The entire table with complete number of rows is provided as supplementary material.

Table 2.4: The GC content and number of variable sites between the two *Plasmodium* species

Gene Pair	Number of Variable sites	P.fal variable GC (%)	P.viv variable GC (%)
1	4247	25	60
2	5816	15	79
3	472	20	71
4	2373	23	68
5	603	13	82
...
...
4281	229	11	81
4282	325	20	63
4283	340	13	80

This table represents only a part of the complete table containing the data of the alignment of the two entire genomes (the same dataset I used for Table 2.2 and 2.3). The data represented here are based on only variable sites in orthologous coding sequences. The entire table with complete number of rows is provided as supplementary material.

In order to find out the direction and the pace of the changes in the GC content in the two genomes, I calculated the GC content of conserved sites and variable sites separately and compared the total GC content to the conserved sites as an extant evidence of the genome composition of the most recent common ancestor between the two species. I repeated this analysis for all three codon positions as the rates of the changes differ between different codon positions.

Table 2.5 summarizes the GC content values at conserved and variable sites in the genome of the two *Plasmodium* species. As it can be clearly seen in the Table 2.5, the average GC content in *P. vivax* orthologous subset is significantly higher than in *P. falciparum*. The difference is almost equal to the difference between the GC content of their total coding sequences (Table 2.1) which shows that the difference between the GC content in these two genomes is not due to a set of species specific genes and runs through the entire genome.

On average, in orthologous pairs of sequences between *P. falciparum* and *P. vivax*, conserved sites compose about 55% of the total sites (5,090,117 out of 9,279,147). This value for variable sites is about 45% (4,189,030 out of 9,279,147). It is shown in Table 2.5 that the GC content value of these sites (i.e. conserved sites), which is 28%, is much closer to this value in all sites (i.e. conserved and variable sites combined) in *P. falciparum* (25%) comparing to the GC content in all sites in *P. vivax* (46%). As one can expect the difference between the GC content value in conserved sites and this value in *P. vivax* is even more pronounced when we take only the variable sites into account (i.e. the difference between the variable sites in *P. vivax* and the conserved sites is equal to 44%).

Table 2.5: The average GC content in conserved and variable sites of coding sequences in two *Plasmodium* species

	Average GC Content (%)	
	<i>P. falciparum</i>	<i>P. vivax</i>
Conserved sites	28	28
Variable sites	19	72
All sites (conserved and variable combined)	25	46

The values are represented in percentages (P -value < 0.001). The dataset used for this table is the same as Tables 2.2 to 2.4.

Figures 2.4 (A, B, and C) represent the GC content values in conserved sites, variables sites, as well as total sites (conserved and variable sites combined) in all codon positions combined. As it was explained earlier and it is clearly shown in this figure as well, the GC content values in *P. falciparum* total sites (i.e. conserved and variable sites combined) (25%) are closer to those values of the only conserved sites (28%). On the other side, this value in *P. vivax* (46%) is far from the conserved sites (Figure 2.4 A and B).

Once we take only the variable sites into account (Figure 2.4.C), these differences are much more pronounced than when we look at the total sites. Thus apparently between these two species, the one species which is diverging from their common ancestor with a faster pace is *P. vivax* even though its GC content value shows a fairly unbiased amount. I will discuss this more in details in the discussion part.

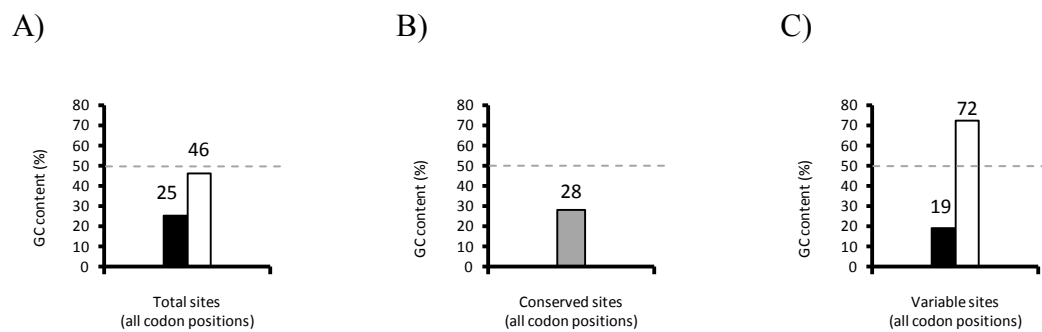


Figure 2.4: GC content of the orthologous coding sequences in *P. falciparum* and *P. vivax*. **(A)** GC content of the total sites (P -value < 0.001) **(B)** GC content of only conserved sites **(C)** GC content of only the variable sites (P -value < 0.001).

I also calculated the same values for the first, the second, and the third codon positions separately in conserved sites, variable sites, and total sites (i.e. conserved sites and variable sites combined) (Figure 2.5). The difference between GC content value of total sites in *P. vivax* and those of conserved sites are the most pronounced when we look at third codon position. This could be expected since the rate of changes is much higher at third codon positions. This difference is as much as 49% when we look at the third codon positions in only variable sites (Figure 2.5.C) and compare it with the conserved sites in this codon position. This will be discussed more in the discussion part.

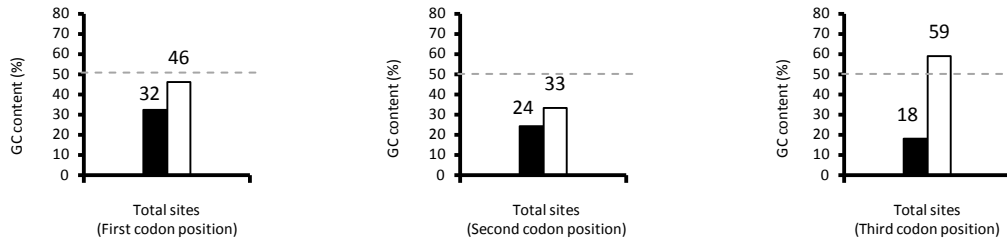
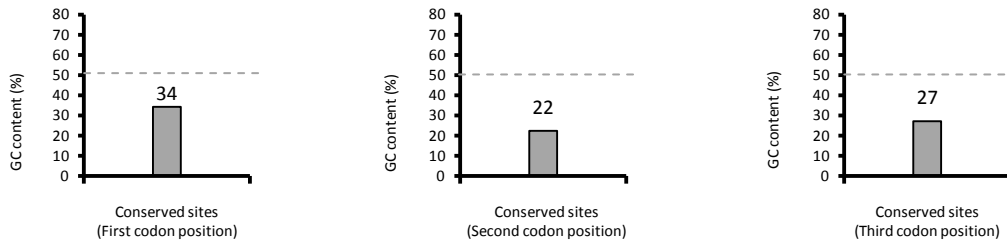
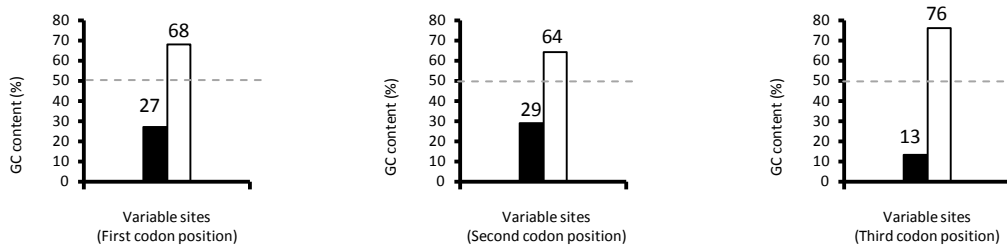
A) Total**B) Conserved sites****C) Variable sites**

Figure 2.5: GC content of the dataset of orthologous coding sequences in *P. falciparum* and *P. vivax* in three codon positions. **(A)** GC content in all the sites (variable and conserved sites) in different codon positions. **(B)** GC content of only conserved sites in different codon positions. **(C)** GC content of only variable sites between the two orthologous subsets of coding sequences in different codon positions. In parts (A) and (C), the black bars represent the data from *P. falciparum* and the white bars represent the data from *P. vivax* (in all cases P -value < 0.001). The dataset used for these histograms is the same dataset used for Figures 2.2.

To find out which one of the nucleotides has a more important role in the observed trends, I calculated the content of each of the nucleotides separately. I did this calculation in each codon position in both conserved and variable sites as well as in total sites (conserved and variable sites combined). Figure 2.6 represents the data for the total sites (conserved and variable sites combined) in all codon positions as well as each of the three codon positions separately.

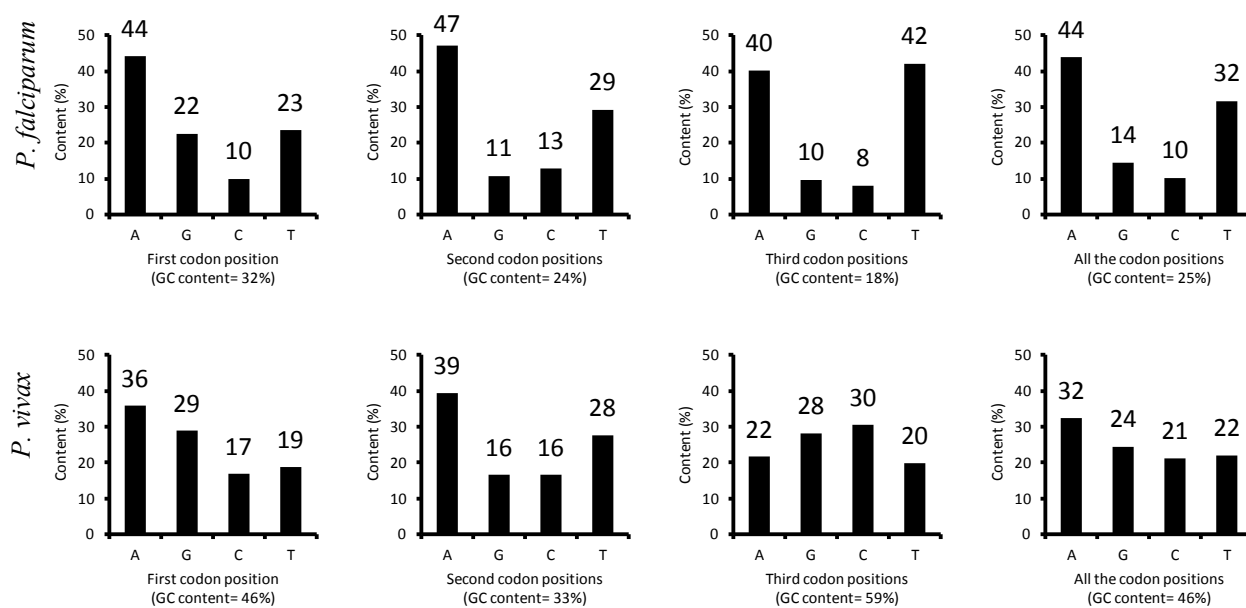


Figure 2.6: Individual nucleotide content of the dataset of orthologous coding sequences in *P. falciparum* and *P. vivax*. The histograms on the top row represent the *P. falciparum* data while those on the bottom row represent the *P. vivax* data.

As it is shown in this figure (Figure 2.6) the differences in GC contents are not distributed evenly between different nucleotides. I also studied the content of each individual nucleotide, in both variable and conserved sites separately. The results of this study are provided as Supplementary material (Figure S.2A, and S.2B), although we do not discuss these data here.

2.4. Discussion

As I explained in Chapter 1, even though it is generally true that close species tend to show similar values in their genomic nucleotide content, but this, in fact, is not always the case and as I showed in my research, in some closely related species there can be a wide difference in their genomic content even within a genus. One very good example for this lack of correlation between the biological relatedness and the nucleotide content of genomes is the difference between the nucleotide content of two *Plasmodium* species causing malaria in human; *P. falciparum* and *P. vivax*. My calculations showed that between these two genomes, there is a large difference in terms of their nuclear genome nucleotide content. It appeared to me that *P. vivax* has maintained a relatively unbiased amount of genomic nucleotide content (GC content = 44%), while the other parasite species shows a very biased low GC value for its nucleotide content. But, I did not stop here and tried to find out what actually is happening in this lineage and especially between these two species since the time of their divergence.

Moreover, the long standing debate on the origin of malaria parasites made me wonder that there might be some unconventional evolutionary events happening in this lineage since diverging from their most recent ancestors. This was the basis of this part of my studies in which I talked about in this chapter. In this part of my research I investigated the bias in nucleotide content and its direction between these two species (i.e. *P. falciparum* and *P. vivax*). Looking at the very low GC content in *P. falciparum* comparing the relatively unbiased content in *P. vivax* one might intuitively conclude that since the divergence of these two species from their most recent common ancestor (MRCA), *P. falciparum* has been losing its GC content while *P. vivax* has maintained the same unbiased value of its GC content. In this chapter however, I show that this, in fact, is not what that has been happening in these two species.

This quite naive interpretation of the current state of the nucleotide contents in the two *Plasmodium* species (i.e. *P. falciparum* losing its GC content after diverging from *P. vivax* while *P. vivax* maintaining its GC content) arises from the idea that when we talk about bias in the nucleotide content of a species, we assume a constant rate and direction of changes in the content. While what I show in this part of my study, indeed, is the fact that this is a very simplistic way of looking at the evolutionary events shaping the GC content of a species' genome. To make it more clear, one can say, considering a common ancestor with a specific GC content for all the species, if the changes in the nucleotide content of a genome were happening in one direction and with a relatively constant rate, how would we be able to see this wide spectrum of GC contents among different species and lineages? This rather simple fact has been mostly neglected when researchers study the GC content behaviour of different species. In my study however, I pinpoint this fact that the bias in the GC content of a specific genome can change back and forth and my results actually show that this changing can happen as fast as we can see the results within the same genus between two close species (i.e. *P. falciparum* and *P. vivax*). This is actually why I call this change in the direction of the bias in nucleotide content “ebb and flow” of the genomic nucleotide content. Basically I show that this changing back and forth of the genomic nucleotide content does not need a very long time to show its effects and it actually can happen as quickly as we even can see its effects between two species of the same genus.

In *Plasmodium* case, I showed that, their common ancestor had a very low GC content and after these two parasites diverged from each other, *P. falciparum* has not changed its nucleotide content as much, but *P. vivax* has started gaining more Guanine and Cytosine content very rapidly (changing its nucleotide content bias in the opposite direction). One might say this could

be because of a possible host switch after the divergence between the two *Plasmodium* species (new environment causing different selective constraints), but my results show that these dramatic changes in *P. vivax* nucleotide content are (at least partly) due to a biased mutation pressure since the majority of the changes found to be happening in the synonymous sites (e.g. third codon positions). Thus one can conclude that the nucleotide content of a genome might go back and forth (referred to ebb and flow in the thesis title) several times through evolution. In this chapter I also show that this ebb and flow of the genomic nucleotide content can represent a different story for each nucleotide.

Chapter 3. Nucleotide bias can distort the usual transition-transversion ratio

3.1. Introduction

In previous chapter (i.e. Chapter 2) I studied the frequency of genomic nucleotide content in two species causing malaria in human; *P. falciparum* and *P. vivax*. I showed that especially in the case of *P. vivax*, we could see a back and forth (or gain and lose) behaviour in terms of its GC content. In the same chapter as well as Chapter 1, I explained that the reason behind studying this behaviour was to have a broader yet more accurate knowledge about the biases (their current state as well as their direction) in the genomic nucleotide contents within this lineage. I also explained that these results could help researchers to improve their methods inferring phylogenetic relationships among different species by correcting the methods themselves or the models of molecular evolution of DNA used by those methods. Moreover, in Chapter 1, I explained that different models of molecular evolution of DNA define the frequency of different nucleotides in each genome as well as the substitution rates between different nucleotides between two genomes. Some of these models take the substitution rates between any two Purine nucleotides as well as between any two Pyrimidines as whole (i.e. transition rates) and also substitution rates between any Purine and any Pyrimidine as another parameter (i.e. transversion rates). These models contribute these two parameters along with the frequencies of different nucleotides in inferring phylogenetic relationships (e.g. Kimura's two parameter model by Kimura M. 1980). There are although some other more general models in that take the rates of all possible substitutions between any two nucleotides into account (e.g. GTR model described by Tavaré S. in 1986).

In Chapter 1, I also explained that there is a bias towards more transitions even though the number of possible transitions is half as great as the number of possible transversions (Figure 3.1). I also explained the reasons behind the usual biased ratio between the transition and transversion rates towards higher transition rates as it was explained by Wakeley (1996). In this chapter however, I show that the bias in nucleotide content can affect this bias between the rates of transitions and transversions.

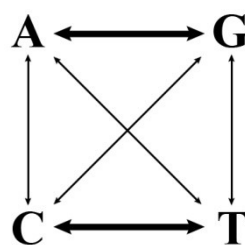


Figure 3.1. Schematic representation of different nucleotide substitutions between four different nucleotides. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

In this chapter, the same as Chapter 2, I study the two human malaria parasites; *P. falciparum* and *P. vivax* as my model organisms. To show the effects of the nucleotide bias on the ts/tv ratio, I compare my results from these two model organisms with other pairs of model organisms with different genomic nucleotide contents. As Brown *et al.* showed in their studies of mitochondrial genomes of five primates (1982), the ts/tv rate ratio in mitochondrial genomes are much more pronounced comparing the nuclear genome. Thus I chose the mitochondrial genome of these two species in order to see the effect of the biased nucleotide content on the ts/tv rate ratio between these two species genomes.

In the same study Brown and his colleagues showed the effect of saturation on the rate ratio of transition over transversion in more diverged species. In order to study the possible substitution saturation in my case, I chose a closer species to one of the model organisms (i.e. *P. falciparum*). This sister species is *P. reichenowi* which has diverged from *P. falciparum* very recently and shares a lot of different biological features with *P. falciparum* (Evans A.G. and Wellems T.E. 2002, Blanquart S. and Gascuel O. 2011). Then I compared my results from *P. falciparum* vs. *P. vivax* comparison with the results from *P. falciparum* vs. *P. reichenowi* comparison in order to see if the substitution saturation has slowed down the rate of substitutions between the former pair since they have diverged long time before the divergence of *P. falciparum* and *P. reichenowi*.

I also compare my results with mitochondrial genomes of a pair of mammals with unbiased nucleotide contents (i.e. Human and Chimpanzee) in order to see if I can see the same rate ratio bias between a pair of unbiased mitochondrial genomes as well or the observed behaviour is due to the nucleotide bias. I also compare my results with the nuclear genomes of *P. falciparum* and *P. vivax* as well as nuclear genomes of human and chimpanzee. The reason I do these further comparisons is the fact that contrary to the biases in mitochondrial genomes of the two *Plasmodium* species which are both towards the same direction, the bias in nuclear genomes of the two *Plasmodium* species, as I showed in the previous chapter, are towards opposite directions. Also I compare my results to the mammalian nuclear genomes to study the *ts/tv* rate ratio in between two unbiased nuclear genomes.

In order to do these comparisons I aligned the aforementioned model organisms' genomic datasets in each pair and built substitution matrices based on the aligned sequences. The methodologies I used in this study are explained more in detail in the Methods section.

3.2. Methods

In order to investigate the *ts/tv* rate ratio between the mitochondrial genomes of the two *Plasmodium* species; *P. falciparum* and *P. vivax*, I extracted the protein and coding sequences of three mitochondrial genes, Cox I, Cox III, and Cyt b from the mentioned *Plasmodium* species from PlasmoDB databank (<http://www.plasmodb.org>). *Plasmodium* species have lost all their protein coding genes in their mitochondrial genomes except the three genes Cox I, Cox III, and Cyt b. The GI number for the mitochondrial complete genome for *P. falciparum* (isolate 3D7) is 31789515 which contains coding sequence data for Cox I, Cox III, and Cyt b (and links to their corresponding protein sequences). The GI number for *P. vivax* (isolate Sal-1) mitochondrial complete genome is 266630826.

Using ClustalW (Larkin M.A. et al 2007), I pairwise aligned each pair of homologous coding sequences in the two *Plasmodium* species together. In order to perform the pairwise alignment, I first aligned their protein sequences and removed the gap columns. Then I assigned the corresponding nucleotide sequences to the alignment files. I used protein alignments to avoid any possible effect of codon degeneracy on the alignment results. Those sites with removed amino acids columns were altered for the corresponding codons to avoid any miscalculation because of any possible frame shift in the assigned nucleotide sequences. Next I concatenated the aligned sequences and produced a single pairwise alignment of two concatenated sequences in which each of the sequences consists of three mitochondrial genes, Cox I, Cox III, and Cyt b with the same number of nucleotides while their gap columns were removed.

Next I calculated the number of identical and non-identical nucleotides between each of the two aligned sequences in each alignment file and built the substitution matrices for each pair of the mitochondrial genomes. The results are shown in Figure 3.2. I repeated this for each of the

three codon positions separately. The results are provided as Supplementary Figure S.3 (C through H).

Using the concatenated sequences I also calculated the content of each of the nucleotides in conserved and variable sites as well as “all sites” (conserved and variable sites combined) (Tables 3.1). I also calculated these values for each pair of orthologous sequences separately in order to investigate the possibility of any artefact arising from a strangely long or short gene with extreme nucleotide content. The results of these analyses are shown as Tables 3.2 through 3.5.

In order to investigate the possibility of saturation, I also needed to compare *P. falciparum* genome with a relatively closer species. As I explained earlier, I chose *P. reichenowi* which is the closest sister species to *P. falciparum* in almost all the phylogenetic studies. Again for this case, I built an aligned dataset of three concatenated mitochondrial genes (Cox I, Cox III, and Cyt b) for both *P. falciparum* and *P. reichenowi* the same way I did for *P. falciparum* and *P. vivax*. Figure 3.3 represents the substitution profile between the mitochondrial genomes of *P. falciparum* and *P. reichenowi*.

Using the same methodology, I also calculated the nucleotide content in conserved and variable sites as well as “all sites” (conserved and variable sites combined) for both *P. falciparum* and *P. reichenowi*. Supplementary Table S.6 represents the data for this calculation. As well, I repeated the same calculation for each pair of orthologous sequences between these two species. Supplementary Tables S.7 through S.10 represent the data in both “all sites” (conserved and variable sites combined) and only variable sites both in actual numbers and percentages.

In order to compare the observed trend of substitutions in *Plasmodium* species with other model organisms without the bias in their nucleotide content, I chose human and chimpanzee as model organisms. These two close mammalian species show fairly normal and unbiased distributions of nucleotide content in their mitochondrial genomes. I built the substitution profile between two mitochondrial genomes of these two species. For this purpose I only used three mitochondrial genes I studied in *Plasmodium* species case (i.e. Cox I, Cox III, and Cyt b) to keep the consistency of my results. I downloaded the protein sequences of these three genes in both human and chimpanzee from GenBank, pairwise aligned them as explained earlier in this section, and assigned the coding sequences to the aligned protein sequences after they were trimmed for gaps and flanking loose ends. Figure 3.4 represents the data obtained from this calculation.

To be able to compare my results with the nuclear genomes, I picked 20 random protein sequences from *P. falciparum* as my reference sequences and found their orthologous sequences in *P. vivax* nuclear genomes (same methodology I used in Chapter 2 in order to find orthologous sequences using protein sequences). Here is the list of the proteins I picked from *P. falciparum* along with their orthologous sequences from *P. vivax*:

MAL7P1.76 and PVX_000615, MAL7P1.77 and PVX_000610, MAL7P1.79 and PVX_000605, MAL7P1.81 and PVX_000590, MAL7P1.82 and PVX_000580, MAL7P1.83 and PVX_000575, MAL7P1.86 and PVX_000565, MAL7P1.87 and PVX_000550, MAL7P1.88 and PVX_000535, MAL7P1.89 and PVX_000530, MAL7P1.91 and PVX_000525, PF07_0066 and PVX_000625, PF07_0067 and PVX_000620, PF07_0068 and PVX_000600, PF07_0069 and PVX_000595, PF07_0070 and PVX_000585, PF07_0071 and PVX_000560, PF07_0072 and PVX_000555, PF07_0073 and PVX_000545, PF07_0074 and PVX_000540. In each pair the

first ID belongs to *P. falciparum* and the second one belongs to the orthologous sequences in *P. vivax*.

However, since I also wanted to compare these results with the genomes of other model organisms such as human, chimpanzee, and *P. reichenowi*, I filtered this list and left out those that did not have orthologous sequence in all of the mentioned species. The final dataset consists of six proteins:

MAL7P1.89 (gi: 296004863) and PVX_000530 (gi: 156094484), PF07_0073 (gi: 124511893) and PVX_000545 (gi: 156094490), PF07_0072 (gi: 124511889) and PVX_000555 (gi: 156094494), PF07_0071 (gi: 124511887) and PVX_000560 (gi: 156094496), MAL7P1.86 (gi: 296004859) and PVX_000565 (gi: 156094498), MAL7P1.81 (gi: 124511875) and PVX_000590 (gi: 156094508). Again, in each pair, the first ID belongs to the coding sequence in *P. falciparum* and the second one belongs to *P. vivax* genome. Numbers in parentheses are the GenBank GI numbers.

Using the same methodology explained for mitochondrial genomes, I built the concatenated orthologous sequence dataset for *Plasmodium* nuclear genes (six nuclear coding sequences), then removed the gaps and calculated the substitution matrices between the two concatenated orthologous sequences. The results of this study are shown in Figure 3.5.

Next in order to compare my results with the nuclear genomes of two vertebrates again I used genomes of human and chimpanzee. To keep the consistency in my results, I used six sequences orthologous to the same set of six coding sequences that I used for *Plasmodium* nuclear genes. The only difference here was that since one pair of the orthologous sequences between human and chimpanzee consisted of a very large stretch of variable sites (comparing the other five pairs)

and the fact that these variable sites showed a significantly different GC contents (80% for human gene and 22% for the chimpanzee gene), I removed this strange pair of sequences to avoid any bias coming from this single pair of sequences in my analysis. Thus I updated my dataset to five remaining pair of orthologous nuclear sequences between human and chimpanzee: 62896524, 37067, 197927451, 15929858, 19068220 for human and for the chimpanzee: 332808314, 345110609, 332816968, 332853013, and 332809722. Then I built the concatenated sequences and removed the gaps the same way I did for the previous datasets. Figure 3.6 represent the substitution profile between human and chimpanzee nuclear genomes.

I also calculated the nucleotide content in both conserved and variable sites as well as total sites (conserved and variable sites combined). Supplementary Table S.1 represents the nucleotide content at different sites in these two genomes. I then calculated the nucleotide content for each of the orthologous pairs (the same as what I did for *Plasmodium* mitochondrial genomes) in both “all sites” (conserved and variable sites combined) and variable sites only. The results of this calculation are represented in Supplementary Tables S.2 through S.5. The GI numbers of the coding sequences used in this study are shown in the caption of Supplementary Table S.2. Data in Tables S.2 and S.4 are represented as actual numbers while these same data are shown in percentages in Tables S.3 and S.5.

3.3. Results:

Figure 3.2 (A and B) represent the substitution profile between two datasets of three concatenated, aligned orthologous mitochondrial genes in *P. falciparum* and *P. vivax* in variable and conserved sites combined in all codon positions.

A)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	891	17	6	122
	G	32	416	3	12
	C	11	4	341	72
	T	111	5	47	1219
	Total	1045	442	397	1425

B)

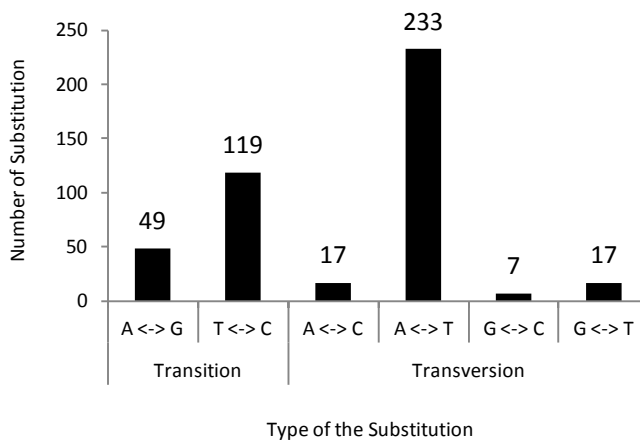


Figure 3.2. Nucleotide substitution profile between the mitochondrial genomes of two *Plasmodium* species; *P. falciparum* and *P. vivax*. **A:** Nucleotide substitution matrix between two datasets. This matrix contains the data for “all sites” (conserved and variable sites combined) in all codon positions represented in actual numbers. **B:** A histogram representing the pattern of different nucleotide substitutions between two datasets.

As it is clearly shown in Figure 3.2 (A and B) the total number of transversions between these two mitochondrial genomes outruns the total number of transitions with a very large margin (i.e. 274 transversions as opposed to 168 transitions). More interestingly even the number of changes between nucleotides Adenine and Thymine alone is larger than the other types of changes between any two nucleotides (both transitions and transversions). The amount of changes between nucleotides Adenine and Thymine between these two genomes is 233 cases which is larger than the total number of transitions combined (168 cases). Even among different types of transversions (274 cases in total), changes between Adenine and Thymine nucleotides compose a very large fraction (85%).

I repeated these steps for each codon position separately (Supplementary Figure S.3) and showed that this inverted bias towards more transversions is only happening at third codon position. Looking at the Supplementary Figures S.3 (C through H), one can notice that in the first and second codon positions, the ratio of transitions over transversions are both greater than one (1.33 for first codon position and 1.06 for the second codon position). Interestingly regardless to the fact that there is no selection pressure on the transitional changes at third codon position (which usually makes the transition bias in this codon position much more pronounced than the other two codon positions), here in this case we can clearly see that not only this bias is not bolder, but in fact it is reversed and the number of transversions outruns the number of transitions with a very large margin (214 transversions comparing to 93 transitions). The *ts/tv* ratio at third codon position in this case is 0.43:1.

More interestingly, when we look at different types of changes at this codon position (i.e. third codon position), we can see that the changes between Adenine and Thymine nucleotides compose a very big fraction among all the different transversion possibilities (201 out of 214

changes, ~ 94%). These amounts are 63% and 29% for the first and second codon positions respectively. Thus one can say that the reversed bias in *ts/tv* ratio is mostly due to the very large number of changes between Adenine and Thymine at the third codon position.

Table 3.1 represents the nucleotide content in conserved sites as well as variable sites and “all sites” (conserved and variable sites combined). Values are represented as percentages. The two mitochondrial genomes show a relatively low GC content (27% for *P. falciparum* as opposed to 25% for *P. vivax*). Although the difference between the GC content values in their variable sites are larger as *P. falciparum* has a 30% GC content while *P. vivax* shows a GC content value of 18%. The GC content of the conserved sites between the two species has a value of 30%.

Table 3.1. Nucleotide content in different sites of three mitochondrial genes in *P. falciparum* and *P. vivax*

	A		G		C		T	
	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>
Total Percentage	32	32	14	13	13	12	42	43
Conserved Percentage	31	31	20	20	10	10	39	39
Variable Percentage	33	35	10	6	20	12	37	46

The top row represents the data in all sites together, the conserved sites only is shown in the middle, and the bottom row represents the data for variable sites only. The data are shown in percentages. The three concatenated genes include Cox I, Cox III, and Cyt b.

In order to rule out any possible artefact in the observed behaviour coming from a specific gene, I calculated the nucleotide content in each of the mitochondrial gene pairs as well. The results of these calculations are shown in Tables 3.2 through 3.5. Table 3.2 represents the nucleotide content data for each pair of orthologous genes in actual numbers along with their

alignment length. Table 3.3 represents the same data in percentage for each of the three pairs of orthologous genes. The last two rows in both tables are the average values and the confidence interval values calculated for significance level (α) = 0.01 (confidence level of 99%). The last column in Table 3.3 represents the data for GC content difference between each pair of the orthologous coding sequences. It is clearly shown that the differences between GC contents of each pair of orthologous genes are not very significant (between 1 to 2 percent).

I repeated these steps for each of the three genes in mitochondrial genomes of the *P. falciparum* and *P. vivax* and showed that the values for the ratio of transition over transversion and also the excess of changes between Adenine and Thymine nucleotides show the same trend as the total genome sequence. The ratio of transition over transversion changes in Cox I, Cox III, and Cyt b are 0.68:1, 0.48:1, and 0.66:1 respectively. Among all different transversional changes, changes between nucleotides Adenine and Thymine compose 89%, 80%, and 85% in Cox I, Cox III, and Cyt b respectively (comparing the 85% when we take the entire genome into account).

Tables 3.4 and 3.5 represent the nucleotide content data of only the variable sites from the same datasets used in Tables 3.1 to 3.3 (i.e. *P. falciparum* and *P. vivax* mitochondrial genomes).

If we normalize the number of variable sites to the length of the alignments in each pair of orthologous genes, we can see that each of the gene pairs shows an almost same proportion of variable sites. The normalized values of variable sites for Cox I, Cox III, and Cyt b are 0.13, 0.16, and 0.13, respectively. The average is equal to 0.14 and the standard deviation for these values is equal to 0.02. The GC content difference between the two species is almost around the same value among all the three individual mitochondrial genes. Tables 3.2 to 3.5 show that each of the three individual mitochondrial genes in both genomes has almost the same characteristic

as the other mitochondrial genes have. In other words these data double check my previous results on *ts/tv* rate ratio to make sure that they are not artefacts arising from a specific gene but it runs through the entire mitochondrial genome between the two species.

Table 3.2. Nucleotide content of three mitochondrial genes in *P. falciparum* and *P. vivax*

	A		G		C		T		Alignment length
	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	
Cox I	435	431	219	207	185	176	592	617	1431
Cox III	248	247	84	88	91	78	327	337	750
Cyt b	353	367	160	147	152	143	463	471	1128
Average	345	348	154	147	143	132	461	475	1103
99% C.I.	537	535	388	341	273	286	759	802	1955

The data shown above represents the actual number of different nucleotides in all the sites (conserved and variable sites together).

Table 3.3. Nucleotide content of three mitochondrial genes in *P. falciparum* and *P. Vivax* (in percentages)

	A		G		C		T		GC (%)		GC diff. (%)
	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	
Cox I	30	30	15	14	13	12	41	43	28	27	-1
Cox III	33	33	11	12	12	10	44	45	23	22	-2
Cyt b	31	33	14	13	13	13	41	42	28	26	-1
Average	32	32	14	13	13	12	42	43	26	25	-2
99% C.I.	8	9	12	8	4	7	8	9	15	14	2

The data shown above represent the nucleotide content in percentages in all the sites (conserved and variable sites together).

Table 3.4. Nucleotide content of the variable sites in three mitochondrial genes in *P. falciparum* and *P. vivax*

	A		G		C		T		Number of variable sites
	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	
Cox I	62	58	20	8	34	25	65	90	181
Cox III	43	42	9	13	24	11	41	51	117
Cyt b	40	54	18	5	29	20	57	65	144
Average	48	51	16	9	29	19	54	69	147
99% C.I.	118	83	58	40	50	70	121	196	319

The data shown above represents the actual number of different nucleotides in variable sites only.

Table 3.5. Nucleotide content of the variable sites in three mitochondrial genes in *P. falciparum* and *P. vivax* (in percentages)

	A		G		C		T		GC (%)		GC diff. (%)
	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	<i>P. falciparum</i>	<i>P. vivax</i>	
Cox I	34	32	11	4	19	14	36	50	30	18	-12
Cox III	37	36	8	11	21	9	35	44	28	21	-8
Cyt b	28	38	13	3	20	14	40	45	33	17	-15
Average	33	35	10	6	20	12	37	46	30	19	-12
99% C.I.	46	28	24	41	9	25	24	32	22	16	38

The data shown above represent the nucleotide content in percentages in variable sites only.

As explained earlier in this chapter, in order to investigate the possibility of saturation (resulting from multiple changes at the same site) affecting the observed behaviour between the mitochondrial genome of *P. falciparum* and *P. vivax*, I needed to compare *P. falciparum* with a closely related species. I chose *P. reichenowi* for this purpose as explained in the methods section.

Figure 3.3 (A and B) represents the nucleotide substitution profile between the mitochondrial genomes of *P. falciparum* and *P. reichenowi*. As I expected, the number of changes are not as large as when I compare the mitochondrial genomes of *P. falciparum* and *P. vivax* since the divergence between the two sister species (i.e. *P. falciparum* and *P. reichenowi*) is a much more recent event and these two species show significantly similar biological features.

A)

		<i>P. reichenowi</i>				
		A	G	C	T	Total
<i>P. falciparum</i>	A	1003	10	8	16	1037
	G	7	454	0	2	463
	C	5	0	399	24	428
	T	16	2	20	1346	1384
	Total	1031	466	427	1388	3312

B)

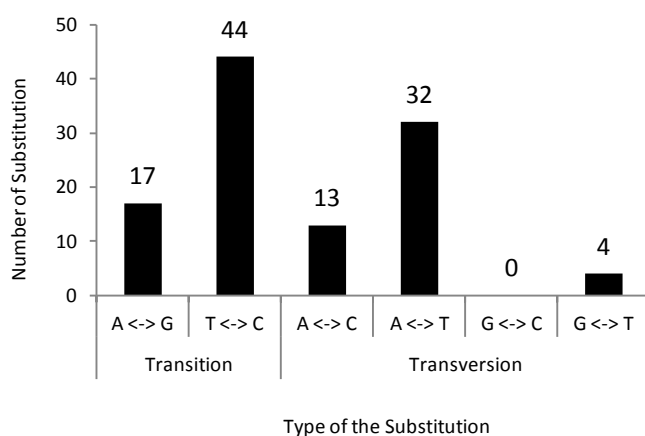


Figure 3.3. Nucleotide substitution profile between mitochondrial genomes of *P. falciparum* and *P. reichenowi*. **A:** Nucleotide substitution matrix between two datasets. This matrix contains the data for “all sites” (conserved and variable sites combined) in all codon positions represented in actual numbers. **B:** A histogram representing the pattern of different nucleotide substitutions between two datasets.

As it is shown in Figure 3.3, the number of transitions between the mitochondrial genome sequences of *P. falciparum* and *P. reichenowi* drops to as low as almost one third of what I observed between *P. falciparum* and *P. vivax* (61 transitional changes as opposed to 168 changes between *P. falciparum* and *P. vivax*). This will make sense if we consider the very shorter time after the divergence between *P. falciparum* and *P. reichenowi* comparing the divergence time between *P. falciparum* and *P. vivax*. On the other side, when we look at the number of transversions between *P. falciparum* and *P. reichenowi* (Figure 3.3) we can clearly see that this number decreases to almost one sixth (49 transversions as opposed to 274 between *P. falciparum* and *P. vivax*). This shows that after the divergence between *P. falciparum* and *P. reichenowi*, these mitochondrial genomes have been accumulating transitional changes with a greater pace than transversions comparing what that has been happening between *P. falciparum* and *P. vivax* after their divergence. Calculating the ts/tv rate ratio shows that after the divergence of *P. falciparum* and *P. reichenowi* this ratio has increased to up to 1.24 (as opposed to 0.61 between *P. falciparum* and *P. vivax*).

Looking at Figures 3.2 (B) and 3.3 (B) it is very clear that the difference in the number of transversions between these two cases is majorly due to the difference between the number of substitutions between the nucleotides Adenine and Thymine (32 changes between *P. falciparum* and *P. reichenowi* as opposed to 233 changes when we look at *P. falciparum* and *P. vivax*). But when we take other substitutions into account these numbers show different behaviours. For example the number of changes between Adenine and Cytosine between mitochondrial genome of *P. falciparum* and *P. vivax* is equal to 17 whereas this number between *P. falciparum* and *P. reichenowi* is 13. This shows that not only the number of changes between Adenine and Thymine between *P. falciparum* and *P. vivax* has not reached to the saturation point (i.e. it

doesn't show a slowdown in the pace its number is increasing) but it is increasing with a greater pace than the other types of changes between these two species. In other words these numbers show that the mutational bias is acting very strongly and has not been slowed down because of saturation through multiple changes happening at the same site. Moreover, even though the selection pressure is stronger on transversions, but apparently in this case the mutational pressure is overriding the selection pressure.

Again in order to show that the observed trend is not an artefact resulted from an extremely long or short gene with a significantly high or low nucleotide content, I repeated this calculation in each pair of the orthologous genes. The results are represented in Supplementary Tables S.6 through S.10 contain the nucleotide content data for each of the three genes in the mitochondrial genome separately.

Supplementary Tables S.7 and S.8 contain the nucleotide content data in all sites (conserved and variable sites combined) in each of the three genes in both actual number and percentages (Table S.7 and S.8 respectively). Supplementary Tables S.9 and S.10 contain the same data but in variable sites only. The last two rows in both tables are the average values and the confidence interval values calculated for significance level (α) = 0.01 (confidence level of 99%).

To compare my results with mitochondrial genomes of other species, I also calculated the nucleotide substitution profile between mitochondrial genomes of human and chimpanzee as two vertebrates with fairly unbiased mitochondrial genomes. As I explained earlier in the methods section, I only took the same three genes as in *Plasmodium* mitochondrial genomes (i.e. Cox I, Cox III, and Cyt b) into account in my calculations. Figure 3.4 represents the nucleotide substitution data between three mitochondrial genes in human and chimpanzee.

A)

		Chimpanzee				
		A	G	C	T	Total
Human	A	901	45	6	2	954
	G	44	456	0	1	501
	C	4	2	978	117	1101
	T	4	0	100	802	906
	Total	953	503	1084	922	3462

B)

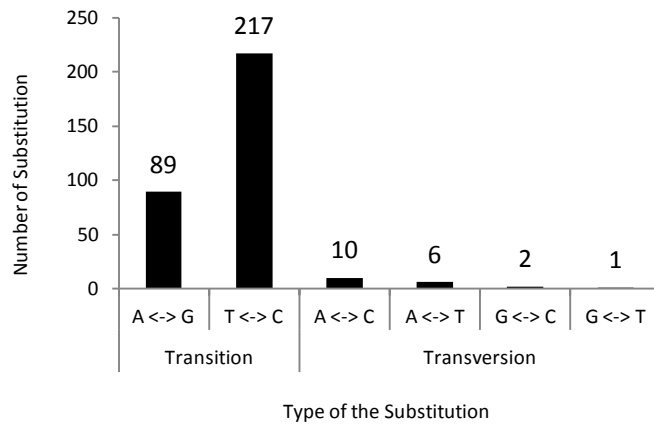


Figure 3.4. Nucleotide substitution profile between two datasets of three concatenated, aligned orthologous mitochondrial genes in human and chimpanzee. **A:** Nucleotide substitution matrix between two datasets. This matrix contains the data for “all sites” (conserved and variable sites combined) in all codon positions represented in actual numbers. **B:** A histogram representing the pattern of different nucleotide substitutions between two datasets.

In order to compare my results with non-mitochondrial genomes and to see if I see the same behaviour in the nuclear genomes of the two *Plasmodium* species (i.e. *P. falciparum* and *P. vivax*) I analyzed a set of six orthologous nuclear genes between *P. falciparum* and *P. vivax* as I explained in the methods section. As shown in Chapter 2, the nuclear genome of these two *Plasmodium* species are both biased but in opposite directions. Once we look at the substitution

profile between these two datasets (Figure 3.5 A and B), we can see that among the variable sites, transitional changes between nucleotides Adenine and Guanine as well as changes between Cytosine and Thymine predominate the other type of changes (2488 changes between Adenine and Guanine and 2387 between Cytosine and Thymine). These numbers are even larger than the number of conserved Guanine and Cytosine nucleotides between the two sequences. The large proportion of variable sites (8470 out of 22350) is consistent with the long divergence time between the two *Plasmodium* species.

A)

		<i>P. vivax</i>				
		A	G	C	T	Total
<i>P. falciparum</i>	A	6115	2209	1196	509	10029
	G	279	2298	257	104	2938
	C	207	215	1411	196	2029
	T	448	659	2191	4056	7354
	Total	7049	5381	5055	4865	22350

B)

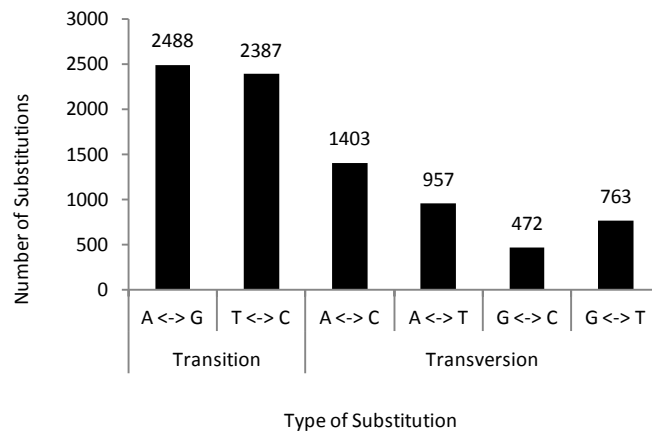


Figure 3.5. Nucleotide substitution profile between two datasets of six concatenated, aligned orthologous nuclear genes in *P. falciparum* and *P. vivax*. **A:** Nucleotide substitution matrix between two datasets. This matrix contains the data for “all sites” (conserved and variable sites combined) in all codon positions represented in actual numbers. **B:** A histogram representing the pattern of different nucleotide substitutions between two datasets.

Contrary to what we saw in their mitochondrial genome however, in these two nuclear datasets, the difference in the GC contents is the result of transitional changes between Adenine and Guanine as well as Thymine and Cytosine (in the latter case, the large number of changes

between Thymine and Cytosine is consistent with the Cytosine de-amination model). Thus apparently since the bias in these two species nuclear genomes are in opposite directions, it hasn't distorted the usual *ts/tv* rate ratio.

As I mentioned in the methods section, I also compared my results to the nuclear genome of two mammalian species. I chose human and chimpanzee as my model organisms and chose same six genes I used analysing *Plasmodium* nuclear genes. Although as I explained earlier in the methods section, I had to remove one of the orthologous pairs because of the very large variable stretch between the two genes and also their strange GC content values.

Figure 3.6 (A and B) represent the substitution profile between two subsets of human and chimpanzee nuclear genomes. Looking at Figure 3.6, obviously the number of variable sites are very small comparing to the conserved sites (they compose only 1% of the datasets). The ratio of transition over transversion is 1.32:1 which follows the usual bias towards more transitions.

A)

		Chimpanzee				
		A	G	C	T	Total
Human	A	4137	32	16	6	4191
	G	22	4752	8	6	4788
	C	9	15	4835	27	4886
	T	8	6	17	3408	3439
	Total	4176	4805	4876	3447	17304

B)

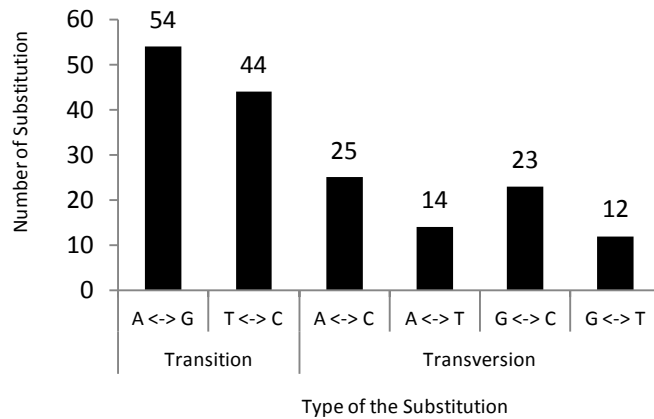


Figure 3.6. Nucleotide substitution profile between two datasets of five concatenated, aligned orthologous nuclear genes in human and chimpanzee. **A:** Nucleotide substitution matrix between two datasets. This matrix contains the data for “all sites” (conserved and variable sites combined) in all codon positions represented in actual numbers. **B:** A histogram representing the pattern of different nucleotide substitutions between two datasets.

Supplementary Tables S.11 to S.15 contain the nucleotide content data (both in actual numbers and in percentages) of each of the gene pairs used in this study separately in all sites (conserved and variable sites combined) as well as in variable sites only. The values in these tables show that the behaviour seen in these two nuclear datasets are not due to a specific gene

with a very long or short length or a significantly high or low GC content and goes through the entire set of the genes almost in the same fashion. Gene identification numbers (GI) for those genes used in this study are listed in the caption of the Supplementary Table S.11 (also mentioned in the methods section).

Looking at the nucleotide substitution profile between nuclear sequences in *P. falciparum* and *P. vivax*, I saw a normal behaviour in terms of transition bias (Figure 3.5). Looking deeper into the abundance of the transitions in this case, one can notice that, interestingly, among changes between Adenine and Guanine, those changes from Adenine to Guanine (from *P. falciparum* to *P. vivax*) compose the larger part of the changes (89%) and among changes between Cytosine and Thymine those changes from Thymine to Cytosine (again from *P. falciparum* to *P. vivax*) compose the majority of the changes (92%). Figure 3.7 represents the difference between opposite directions of transitional substitutions between *P. falciparum* and *P. vivax* nuclear datasets.

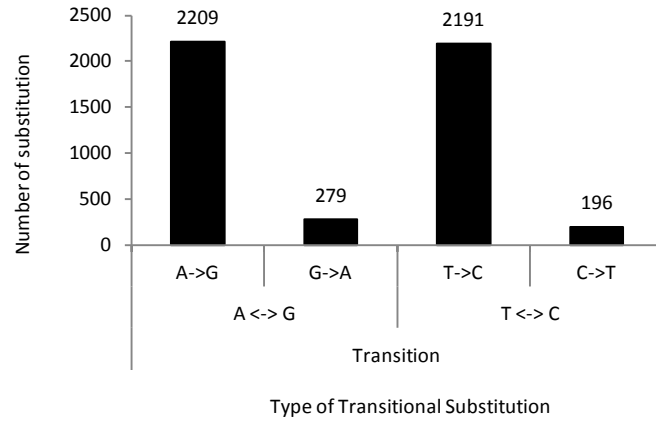


Figure 3.7: The observed pattern of transitional changes between two nucleotides from the same class, between two sets of concatenated, aligned orthologous nuclear genes in *P. falciparum* and *P. vivax*.

In the nucleotide change $X \leftrightarrow Y$, “X” (the first nucleotide) refers to a nucleotide from *P. falciparum* nuclear genome and “Y” (the second nucleotide) refers to a nucleotide from *P. vivax* nuclear genome. Also $X \rightarrow Y$ represents changes from nucleotide “X” from *P. falciparum* nuclear genome nucleotide “Y” in *P. vivax* nuclear genome.

In this case we see even though the bias in opposite direction in these two nuclear genomes does not distort the usual ts/tv rate ratio, but the bias still shows its effect on the evenness of the distribution of the changes between two nucleotides in either opposite direction. In this case, we see in the two different transitional changes ($A \leftrightarrow G$ and $C \leftrightarrow T$), the number of $A \rightarrow G$ changes are not the same as the number of $G \rightarrow A$ changes and the number of $C \rightarrow T$ changes differs from the number of $T \rightarrow C$ changes.

3.4. Discussion:

Other than the frequency of each nucleotide, another aspect of the genomes, that can affect the molecular models of DNA evolution, is the substitution rate between different nucleotides. However, in this part of my study I showed that even this aspect of genomes can be affected by the nucleotide content of a genome. The results of this part of my research clearly show that the bias in nucleotide content can affect the rate ratio of the transitions and transversions. I showed that if there is a bias in the nucleotide content of the two species, and more importantly, if the bias in both genomes is in the same direction, (e.g. the bias in mitochondrial genomes of the *P. falciparum* and *P. vivax*), the usual transition transversion rate ratio can be distorted (Figure 3.2). But if the bias in the two genomes are in the opposite directions, as we saw in the case of nuclear genomes of the two same species, this distortion will not happen and we will witness the usual transition transversion in their substitution profile (Figure 3.5).

In my study I showed that the rate of transversional changes between Adenine and Thymine exceeds its expected value with a very large margin and overrides all the other transitional and transversional substitutions combined (Figure 3.5). This very large number of Adenine to Thymine transversional changes is the consequence of the very low GC content of both genomes. This should be yet another point to keep in mind when designing or modifying a molecular model for the evolution of DNA in which can be used to infer the phylogenetic relationship within this lineage.

Another fact I unveiled in this study is about the saturation in the nucleotide substitution between these two species because of multiple changes in one site. One can simply show saturation by the decrease in the percentage sequence difference in sites subject to transitions over this value in those sites subject to transversions. In this study however, I show that, in fact,

this is not always the case. In this study I showed that even though the ratio of the percentage sequence difference in those sites subject to transitions over this value in those sites subject to transversion decreases between mitochondrial genomes of two *Plasmodium* species, this decrease is not a sign of saturation in transitional changes. I showed that the very large number of changes between Adenine and Thymine between *P. falciparum* and *P. vivax* (transversion) is, in fact, the main reason that the rate ratio of transition over transversion decreases through a dramatic increase in the number of transversions (as opposed to a decrease in the rate of transitional changes which is the basis of saturation). This shows that just a decrease in the ratio of transition over transversion rate by itself is not sufficient to call if transitional substitutions between two nucleotides have reached a saturation point.

Table 3.6 summarizes the transition/transversion rate ratio data in different cases I studied in this chapter.

Table 3.6: The *ts/tv* rate ratio between different genomes.

Genome 1	GC content	Genome 2	GC content	<i>ts/tv</i> rate ratio
<i>P. falciparum</i> mitochondrial	27%	<i>P. vivax</i> mitochondrial	25%	0.6:1
<i>P. falciparum</i> mitochondrial	27%	<i>P. reichenowi</i> mitochondrial	27%	1.2:1
Human mitochondrial	46%	Chimpanzee mitochondrial	46%	16.1:1
<i>P. falciparum</i> nuclear	22%	<i>P. vivax</i> nuclear	47%	1.4:1
Human nuclear	56%	Chimpanzee nuclear	56%	1.3:1

In this table, all the rate ratios are normalized to one.

Another point I unveiled in this part of my study is the possible unevenness of opposite substitutions within a class (i.e. transition or transversion). I showed that this unevenness is

correlated with the difference between nucleotide contents of the two genomes. For example, between nuclear genomes of *P. falciparum* and *P. vivax*, since the GC content of the *P. falciparum* genome is much lower than in *P. vivax*, in two transitional changes, the one between Adenine and Guanine, from *P. falciparum* to *P. vivax* the majority of changes are those from Adenine to Guanine while in changes between Cytosine and Thymine, from *P. falciparum* to *P. vivax*, those from Thymine to Cytosine compose the majority of the changes.

In general, in this part of my study I showed that in the case there is bias in the genomes nucleotide composition, the rates of changes between different nucleotides can be strongly affected by this bias and will deviate from expected values based on Kimura's two parameter model. Especially, if the biases in two genomes nucleotide content are in the same direction it can distort the usual ts/tv rate ratio.

Chapter 4. The relationship between gene length and nucleotide content

4.1. Introduction

In Chapter 1, I listed a series of theories explaining the very significant heterogeneity both between and within genomes. I also mentioned that one can divide these theories into two groups based on the main evolutionary force having the major effect in shaping the variation in nucleotide content. One of these groups suggests natural selection as the major player shaping the heterogeneities and the other one takes biased mutation responsible for the variation. In the literature the former is usually referred to as “selectionist” model and the latter one is referred to as “neutralist” or “mutationalist” model. The selectionist model, which is mostly supported by Bernardi and his colleagues (Bernardi G. and Bernardi G. 1985; Bernardi G. 1993), argues that the higher body temperature in warm-blooded animals (homeotherms) can select against low GC (i.e. high AT genomes) for higher GC content genome. Bernardi explains that this selection can happen both in nucleotide level (because of the difference in the number of hydrogen bonds between Adenine and Thymine and this number between Cytosine and Guanine) as well as in amino acid level. He explains (1986) that the ratio of (Alanine + Arginine) / (Lysine + Serine) increases from GC poor to GC rich genes (Alanine and Arginine mostly contribute in the thermostability of the proteins). Nonetheless, this theory (genomic stability) has its own critics. I mentioned in Chapter 1 that Chojnowski (2007) and Fryxell and Zuckerkandl (2000) showed that, in fact, higher temperature can increase the rate of Cytosine deamination feedback loop which leads to a decrease in the GC content of a genome. Fryxell and Zuckerkandl suggest the thermal mutationalist model (which is a neutralist model) and describes a bolder effect of biased mutation than selection in shaping the heterogeneity in nucleotide contents of a genome.

As explained in Chapter 1, some of the earlier studies in our laboratory lead me to the idea that both biased mutation and selection might have major effects on the genomic make up. Wang *et al.* (2004) compared the genome content of homologous coding sequences from rice and *A. thaliana*. They found that homologous genes between *A. thaliana* and rice show different distributions of their nucleotide content. While *A. thaliana* genome shows a uniform, unimodal distribution (showing only one peak in the probability distribution), the homologous genes of rice, show a broader, bimodal pattern. They showed that there is a subset of genes in rice genome which have become highly GC-rich since the divergence from *A. thaliana*, but in the meanwhile some other genes in this genome have maintained their GC content close to their homologs in *A. thaliana*. To put their findings in simple words, what they found was that a subset of the ancestral genes between these two species (i.e. homologous genes that they share with their common ancestors) has undergone a series of changes that lead to a significant change in their GC content. Now the question is what about this subset of genes was different from the rest of their ancestral genome that made them to undergo these changes while the other part of the ancestral genome did not. Wang *et al.* (2004) in fact had found the answer to this question in the same study but did not relate it to the question since this question was not their main concern. What they had found in their study which led me to develop my hypothesis was the fact that the GC-rich rice genes (i.e. biased genes) were on average significantly shorter than those genes which their GC content had not changed since the divergence from *A. thaliana*. Basically this finding about the differences in lengths is the main difference I was looking for between the two subsets of genes mentioned earlier.

In order to put my hypothesis into test, I needed another model organism. For my model organism I needed to pick an animal with a low GC content and also a genome with a bimodal

distribution of GC content so I would be able to answer the three questions raised (mentioned in Chapter 1); whether the negative correlation between GC content and average length is specific to the genomes of flowering plants, or if it can be seen in other genomes as well, the question that if the negative correlation is between gene length and GC content *per se* or with the degree of nucleotide bias, and the third question about the nature of the evolutionary forces affecting this specific subset of genes. I chose honeybee as my model organism since it meets all the criteria I was looking for. It is an animal with a low GC content genome with a bimodal distribution for its GC content (i.e. a subset of its genes show very low GC content values while another subset of genes show relatively higher GC contents). I divided these genes into two subsets of low and high GC content genes and calculated their average length. This will be explained more in details in the Methods section.

Our results showed that in honeybee genome which is an AT biased genome; there is a negative correlation between the average length of the genes and the Adenine and Thymine content of the genes (in other words there is a positive correlation between the average length of the genes and their GC content). This, at first, seems to contradict with what Wang *et al.* (2004) had found in their study in rice genome. But at second thought, taking the bias in nucleotide content into account, in a high GC genome such as rice, the biased genes are those with higher amounts of GC however in a low GC genome such as honeybee, the biased genes should be considered those with high AT content (i.e. low GC content). Thus in both rice and honeybee genomes, what we see is a negative correlation between the average gene length and the degree of bias in nucleotide content from an unbiased value (i.e. 50%). Thus apparently the shorter the genes are in both genomes the faster they can accumulate changes through the course of evolution. Based on the results in this study and Wang's study, I will come up with the

mechanism that might be behind this negative correlation. We will see that based on my suggesting mechanism, both biased mutational forces and selection have roles in shaping the variation in genomic nucleotide content. My suggesting mechanism is, in fact, based on a well known mechanism called “Background Selection” which has been studied and explained before by many researchers such as Brian Charlesworth (1993, 1995).

Background selection model, which states that deleterious mutations can affect the evolutionary fate of closely linked sites, can somehow be considered as an opposite mechanism to selective sweep (Nurminsky D.I. *et al.* 1998) in which a beneficial mutation can raise the chance of fixation of its neighboring linked sites. Both models however are based on the hitchhiking effect explained by John Maynard Smith and John Haigh in 1974. In both models the genetic diversity of the neighboring sites will reduce but in background selection model this is due to a harmful mutation happening in a site and elimination of the site itself along with its neighbors from the population, however, in selective sweep a beneficial mutation at a site helps to fix an allele in a population along with its neighboring linked sites. I will discuss my results and the background selection model in details in the discussion section.

4.2. Methods

In order to study the genome of honeybee I downloaded 6551 honeybee coding sequences from *BeeBase* database (<http://www.beebase.org/>) (Munoz-Torres *et al.* 2011) alongside with their corresponding proteins, introns and total gene sequences. Next I developed a program to calculate the GC content of the nucleic acid sequences alongside with their length. Also I calculated the amino acid content of the corresponding protein sequences. I double checked my sequence data with the data obtained from *NCBI genome database* (<http://www.ncbi.nlm.nih.gov/genome/>). Having the coding sequence, whole gene sequence, and introns data in hand, first I calculated the GC content distribution in all these three datasets. I used the bin size of one percent and counted the number of genes with GC content within the one percent bin and produced the distribution graphs for all three datasets. The results are shown in Figure 4.1.

To study the correlation between the GC content of the honeybee coding sequences and their length, I sorted the entire set of coding sequences of honeybee based on their GC content and then divided them into two groups of same number (one groups with lower GC content and the other groups those sequences with higher GC content). Then I calculated the average length of each group. The results are shown in Figure 4.2.

To study this correlation in a finer fashion and also to avoid the impact of potentially very long or short genes in either of the subsets on the results, I divided the sequences into five groups of same number of sequences (after sorting them based on their GC content) and then calculated the average GC content and average length for each group. The results are shown in Figure 4.3(A).

I repeated this correlation study for total genes and also intron sequences and compared the GC content of the coding sequence of each gene with the average length of its introns (all the introns combined) and the total gene sequence. The results of this calculation are presented in Figures 4.3(B) and 4.3(C). As it is shown in these figures, an increase in GC content of the coding sequences is consistent with increase in average length of coding sequence as well as introns and total gene sequences, although this increase is considerably more pronounced in introns and total genes than in the coding sequences.

In order to compare the rate of accumulation of the changes in different codon positions, I developed a program to calculate the GC content of each of the three codon positions in the coding sequences and compared the average results in each of the aforementioned five groups with the GC content of coding sequences in the those groups (Figure 4.4).

In order to study the effect of GC content on the amino acid composition of the proteins in honeybee, I calculated the amino acid composition of the proteins corresponding to the genes in each of the five groups mentioned before. The results are shown in Figure 4.5.

Next, to compare my results with those from rice genome I downloaded 58058 rice coding sequences from *EnsemblPlants* (a part of the Ensembl Genomes, Kersey et.al. 2009; <http://plants.ensembl.org/biomart/martview/>). Using the Perl program I developed before, I counted the content of each nucleotide in my sequence dataset and calculated the distribution of the GC content in this genome. Figure 4.6 represents the distribution of the GC content in the genome of rice compared with this distribution in honeybee.

Next I needed to compare the correlation between average gene length and the nucleotide content of the genes in these two genomes. For this purpose I first sorted the coding sequences of

rice based on their GC content and then divided them into two groups of the same number as I did for honeybee genome. Then I calculated the average length of the coding sequences in each group. The results of this calculation along side with the results from honeybee genomes are represented in Table 4.1.

In order to find out what it is special about those genes in which have been able to accumulate the most of the changes I needed to study them separately from the rest of the rice genome. For this purpose I first removed all the sequences with GC lower than 60% from rice genome dataset. The remaining subset of coding sequences included 15871 highly biased coding sequences with GC content greater than 60%. I used this subset of sequences in the comparison study with honeybee coding sequences.

Next, using the standalone version of BLAST program (Altschul *et al.* 1990), I pairwise aligned all the honeybee coding sequences against the high GC subset of coding sequences in rice (GC content higher than 60%); I used 1E-20 expectation value as cut off value for the alignment. Then I compared the distribution of the GC content of the obtained subset of coding sequences in honeybee including 282 coding sequences against the total set of coding sequences in this species. I then compared the GC content distribution in the 282 obtained coding sequences and compared this distribution with the total honeybee genome dataset. The result of this comparison is shown in Figure 4.8.

4.3. Results

Figure 4.1 represents the distribution of GC content in honeybee genome. In this figure we clearly can see the bimodal distribution of the GC content in total gene sequences, coding sequences, as well as introns. Among these three datasets, we can clearly see that while GC contents in introns show relatively lower amounts ranging from 7% to 56%, of GC content in coding sequences are distributed more towards relatively higher values ranging from 21% to 71%.

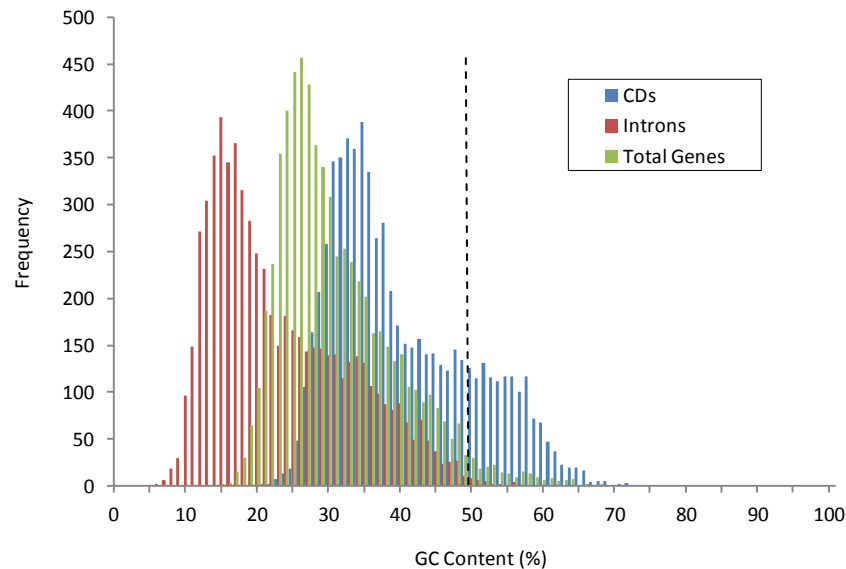


Figure 4.1. GC content distribution in honeybee genome.

GC content of the total gene sequences however are distributed in a relatively broader range between 11% and 65% and the major chunk of this distribution is located between 25% and 35% (the maximum data point of this distribution shows 497 genes with GC content values between 25% and 26%). The maximum points for introns shows 393 sequences with GC content between

14% and 15% and for coding sequences 388 sequences with GC content values between 34% and 35%.

These results show a pronounced bimodal distribution in the GC content of honeybee genes in a way that we can clearly divide the genes in this genome to those who have shifted their GC content to lower values while the other subset of genes have maintained a relatively unbiased GC content values. This was important to me since as I explained in the introduction section of this same chapter I was looking forward to study the causes behind the more accumulation of changes in a specific set of genes through the course of evolution. This clear boundary between these subsets of genes makes it easier for me to compare different aspects of them in order to find the causes that might play major roles in shifting their nucleotide content.

As explained in Methods section, I divided the entire set of coding sequences in honeybee into two halves with the same number of sequences (each containing 3275 coding sequences) after I sorted them based on their GC contents. Next I calculated the average length of the coding sequences in each subset. Figure 4.2 represents the results of this comparison.

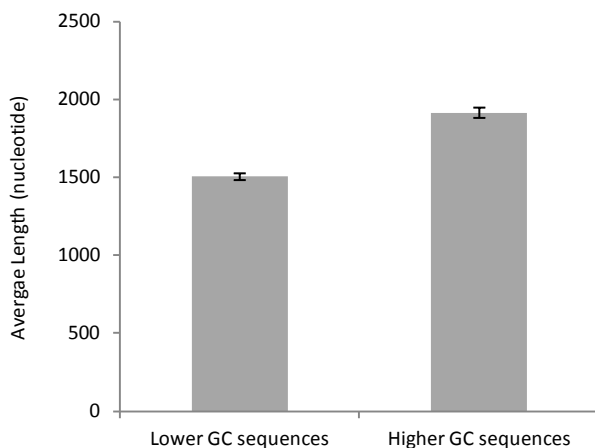


Figure 4.2. Average length of coding sequences in two groups of honeybee genes with lower and higher GC contents. The bars indicate the average length on coding sequences in nucleotides. The error bars are the standard errors ($P\text{-value} < 0.001$).

As it is shown in this figure, the subset of genes with lower values for their GC content are significantly shorter on average than those genes with higher amounts of GC content.

To study this trend in a finer fashion, as explained in the methods section, I divided the total dataset into five groups based on their GC content and again calculated their average length and average GC content values. I repeated this study for introns and total gene sequences as well. Figure 4.3 (A, B, and C) represent the results of these calculations. It is shown in this figure that in all three groups (i.e. coding sequences, introns, and total gene sequences) the average length of the sequence increases as GC content of the coding sequence increases. This clearly shows that this feature runs through the entire genome and is not due to a specific groups of sequences with odd length or GC content values. Since introns compose major part of genes, the graphs for introns and total genes almost look alike in both Y axis scale and the trend (Figure 4.3 B and C). Although we see an increase in the average length of the coding sequences while we move towards higher values of average GC content, but this increase is considerably more pronounced

in introns (and as a result in total gene sequences). The difference between average length of the coding sequences with the lowest GC content and the highest GC content is 590 nucleotides while this difference in introns equals to 23815 nucleotides. For total genes the average length difference between these two groups is equal to 24200. This clearly shows that this difference in total gene sequences is strongly due to the differences in intron length.

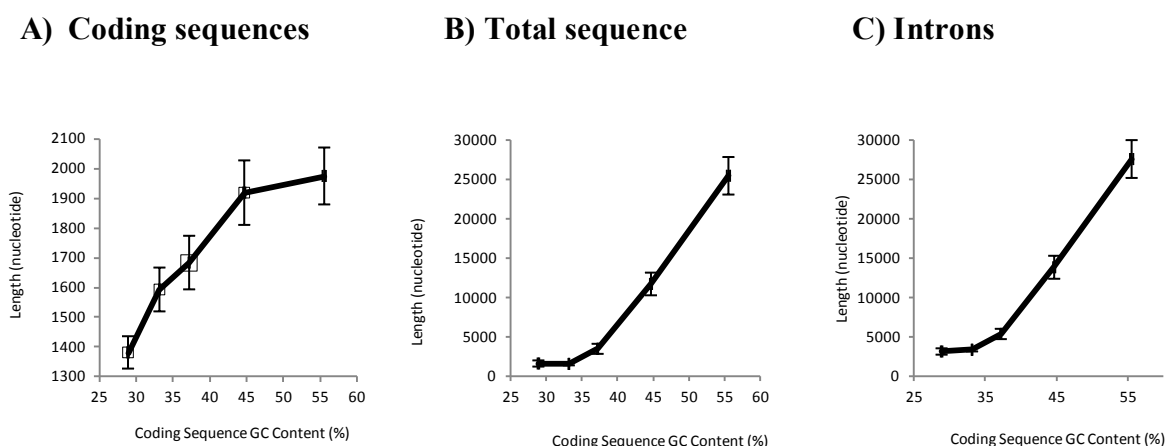


Figure 4.3. The relationship between GC content of the coding sequences with average length of (A) coding sequences (B) total gene sequences and (C) introns (all introns of a gene combined) in honeybee. The confidence intervals are calculated for alpha (significance level)=0.05 (i.e. the confidence level is 95%) for all three graphs.

Looking closer to Figure 4.3 (A), one can realize that the pace of increase in the average length of the coding sequences is slowing down when we move towards highest amounts of GC contents. To understand the causes behind this slowdown, I calculated the proportion of coding sequences in total genes in the same five groups. The results are shown as supplementary Figure S.9. In this figure the higher we move on the axis representing the GC content, the proportion of the gene length occupied by coding sequences decreases. This can explain the slowdown in the changes in the coding sequence length when we move towards the higher values of GC. Of

course one might expect that when we get to the extreme point when the proportion of the coding sequences reaches zero, an increase in GC content will not affect the average length of the coding sequences.

In order to show the possibility of directional mutation pressure in shaping the heterogeneity in the GC content values among honeybee genes, I used the same idea explained in Chapter 2 when I compared the rate of changes between different codon positions. Using the same five groups (sorted and divided based on their GC content) I calculated the GC content in all three codon positions. Figure 4.4 represents the results of this calculation. As one might expect, when there is a mutational bias, the synonymous sites should be able to accumulate more changes than the other two codon positions since they are less sensitive to the selection.

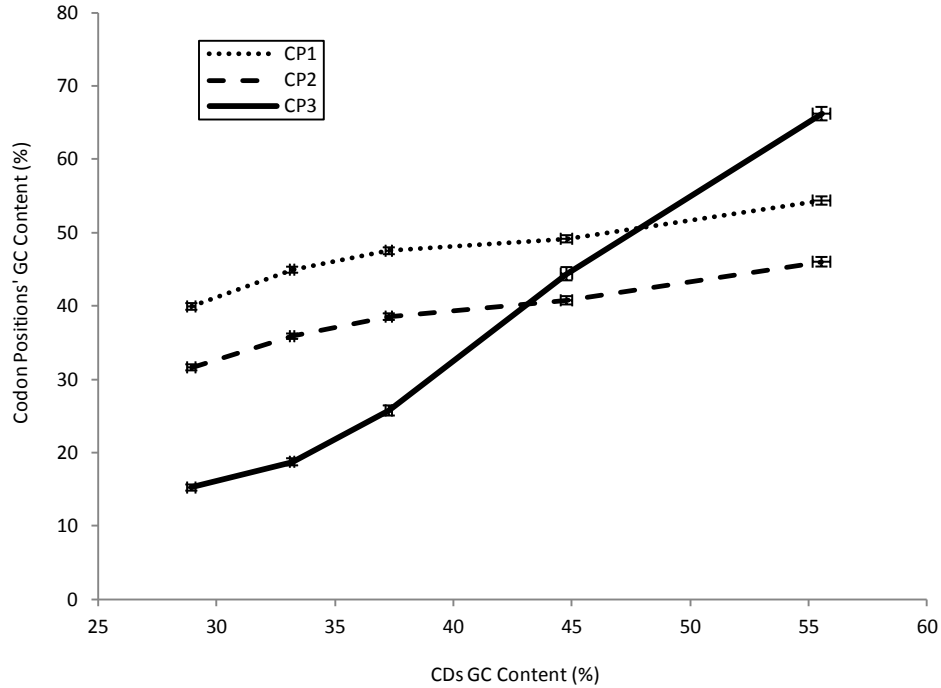


Figure 4.4. The changes in GC content of each codon positions compared to the changes in the GC content of the coding sequences in honeybee. The dotted line represents the changes in the GC content at first codon position, while the dashed line represents the changes in the second codon position, and the straight line shows the changes at third codon position. The bars indicate the the confidence intervals calculated for alpha (significance level) = 0.01 (the confidence level is 99%).

Looking at Figure 4.4 we can clearly see that third codon position shows the highest rate of changes comparing the other two codon positions. This indeed shows that whatever combination of evolutionary forces is behind the observed variation in the genomic nucleotide content in honeybee genome, biased mutational pressure has its own role in shaping it.

Figure 4.5 shows the changes in the proportion of the content of four different groups of amino acids in proteins corresponding to the aforementioned five subsets of coding sequences (which were sorted and divided based on their GC content). The amino acids are grouped based

on the nucleotide composition of their codons (previously explained in Chapter 1 and also mentioned in the caption of Figure 4.5).

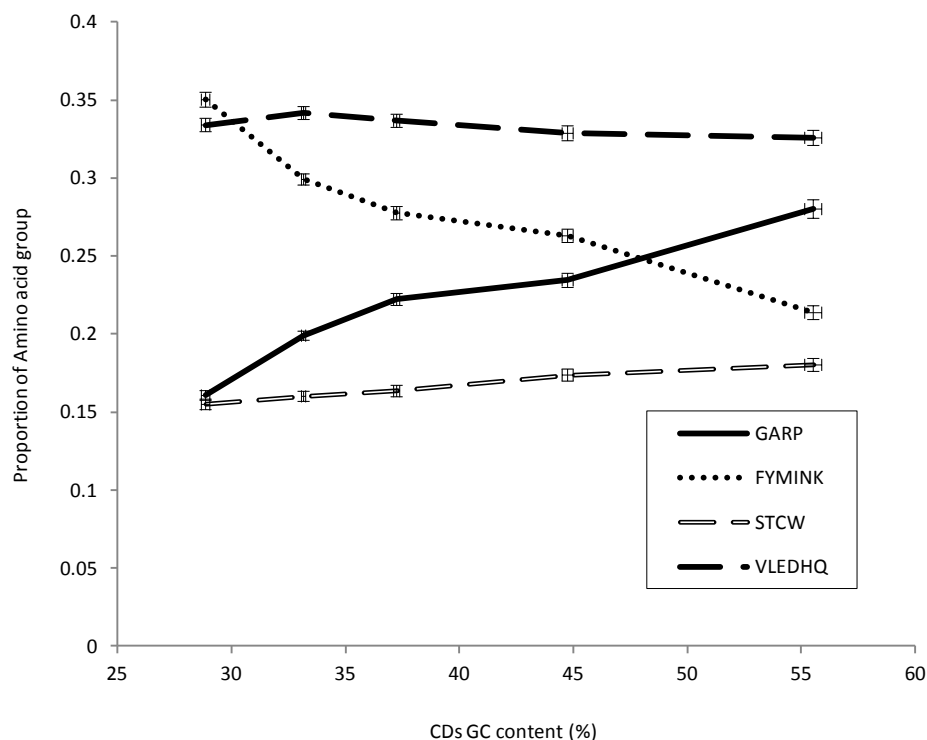


Figure 4.5. Proportion of different groups of amino acids vs. GC content of their corresponding coding sequences in honeybee. GARP represents the amino acids Glycine, Alanine, Arginine, and Proline. FYMINK represents amino acids Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine, and Lysine. STCW represents amino acids Serine, Threonine, Cysteine, and Tryptophan. VLEDHQ represents Valine, Leucine, Glutamic acid, Aspartic acid, Histidine, and Glutamine. The confidence intervals are calculated for alpha (significance level) = 0.001 (the confidence level is 99.9%).

This figure shows that once we move towards those groups of proteins with higher GC content in their corresponding coding sequences (X-axis), the proportion of the amino acids with G/C in

their first two codon positions (i.e. GARP amino acids) increases, while other amino acids with more A/T nucleotides in their first two codon positions (FYMINK) show a decrease in their proportions. The other two groups with unbiased nucleotide composition in their first two codon positions (STCW and VLEDHQ) do not show a significant change in their proportions. This can be interpreted as either the amino acid composition acting as a selective constraint shaping a specific nucleotide composition or, on the other side, the effect of nucleotide composition on the amino acid content in the proteome of honeybee.

To compare the honeybee results with those from rice genome, in order to find inclusive answers for the questions mentioned in the introduction of this chapter, I repeated the same calculations for the homologous subsets of honeybee and rice genomes and compared the results from these two genomes. I first calculated the GC content distribution in these two subsets of genes in both genomes. Figure 4.6 represents the results of this calculation.

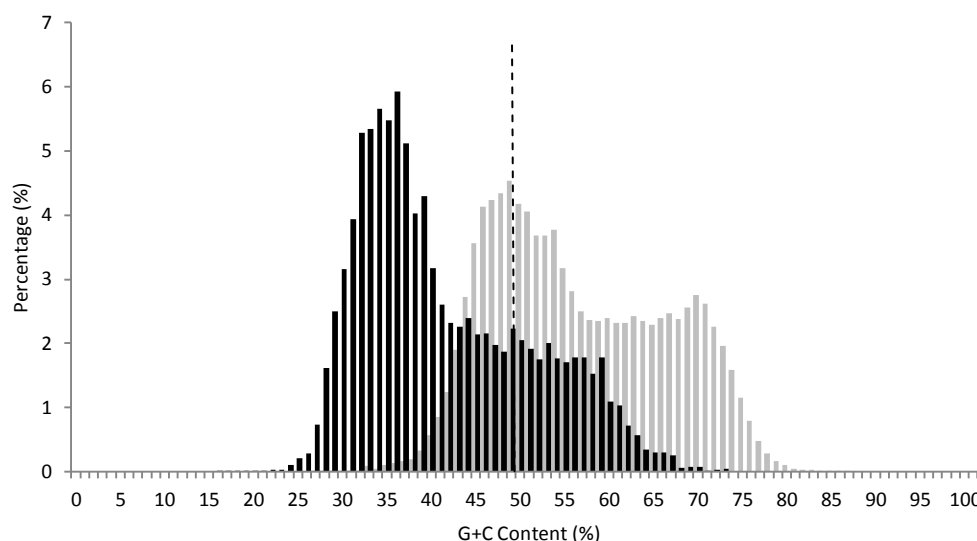


Figure 4.6. The distribution of GC content in honeybee (darker bars) and rice (lighter bars) coding sequences. The darker lines represent honeybee data while the lighter ones represent the rice data.

As it is clearly shown in this figure, both subsets of genes from rice and honeybee genomes represent bimodal distributions in their GC content but in two different directions; the distribution of the GC contents in the subset of genes in honeybee genome is more towards low values of GC while this is the opposite in the rice genome.

I also divided the total set of rice genes into two subsets of the same number after I sorted them based on their GC content (the same I did for honeybee genome). The results of the comparison of the average length and GC contents in both lower GC and higher GC groups for both honeybee and rice genomes are shown in Table 4.1 and Figure 4.7. These results are shown in a graph later in the Chapter 5 (Discussion) in Figure 5.9.

Table 4.1. Average length of coding sequences in two groups of genes in honeybee and rice.

	Honeybee		Rice	
	GC Content	Length	GC Content	Length
Low GC genes	32 ± 0.1	1504 ± 43	46 ± 0.05	1902 ± 18
High GC genes	48 ± 0.3	1914 ± 66	62 ± 0.1	1238 ± 15

Sequences are sorted based on their GC content and then divided into two groups of same number of sequences, one with lower GC contents and those sequences with higher GC contents. The confidence intervals are calculated for alpha (significance level) = 0.05 (i.e. the confidence level is 95%).

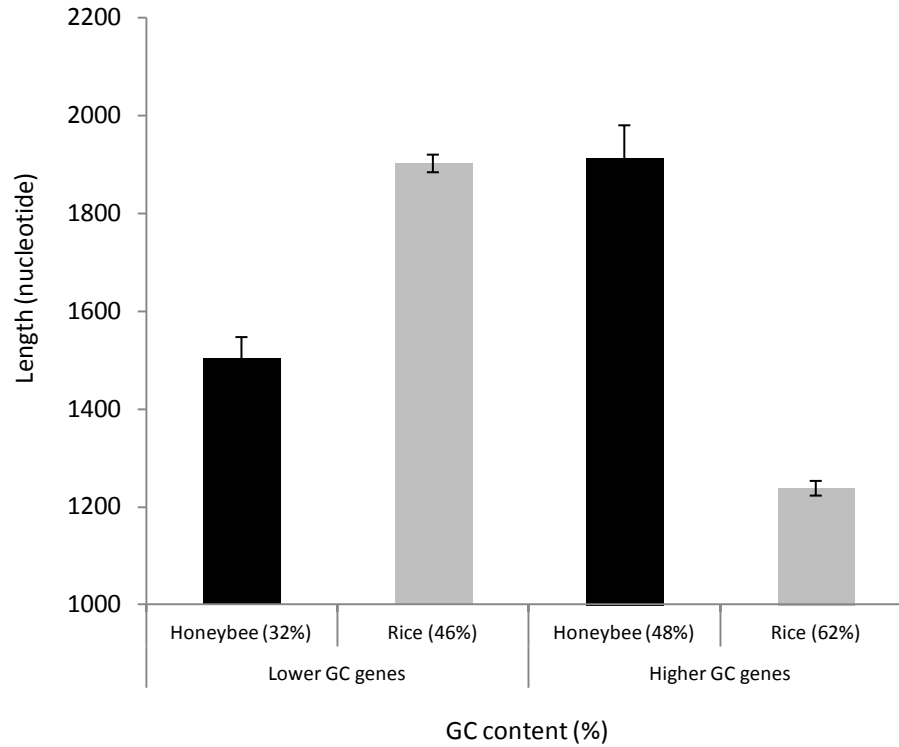


Figure 4.7. Average length of coding sequences in two groups of genes in honeybee and rice. Darker bars indicate the average length of the coding sequences in two honeybee datasets. The lighter bars represent the average length of the coding sequences in two subsets of rice genes. The error bars indicate the standard error values.

It is very clear in Table 4.1 (as well as Figures 4.7 and 5.9) that the subset of genes with a more biased amount of GC content are on average significantly shorter in length than the other subset of genes with GC content values around the unbiased values (48% in honeybee genome and 46% in rice genome).

In order to investigate what has happened in terms of nucleotide content to the sequences homologous to very biased coding sequences in the rice genome, I first filtered out the rice genome for those coding sequences with GC content values greater than 60%. Then I pairwise

aligned these sequences against the total genome sequences of honeybee. Figure 4.8 shows the distribution of the GC content of the resulting subset of sequences (honeybee's coding sequences homologous to rice high GC coding sequences) along with the GC content distribution of honeybee total set of coding sequences.

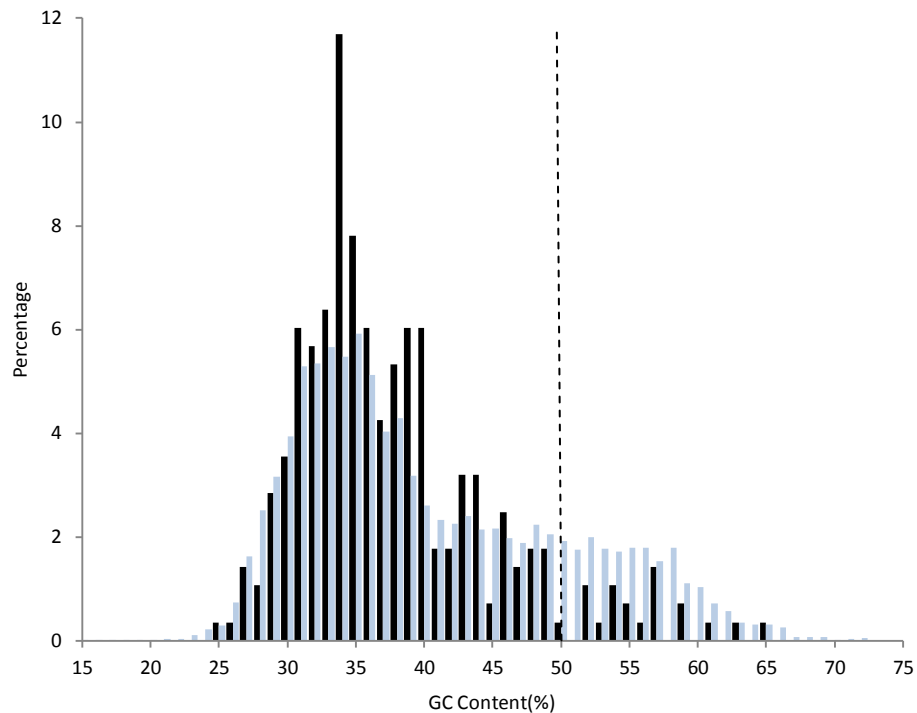


Figure 4.8. GC content distribution in two subsets of coding sequences in honeybee

The lighter lines represent the total coding sequence dataset while the darker lines represent the subset of coding sequences which are homologous to the high GC content subset of rice genes.

In Figure 4.8 it is clear that this subset of honeybee coding sequences that are homologous to very high GC (and on average shorter) rice genes are more leaned towards lower GC contents comparing the original dataset. We can see this in a way that those genes from this subset of data that have GC contents higher than 50% (darker bars on the right side of the dashed line) show

smaller percentages (7%) comparing the original dataset with GC content higher than 50% (lighter bars on the right side of the histogram) that compose 21% of the genes. It is as if this subset of sequences is missing a peak of high GC sequences). These values for genes with GC content less than 50% for the total genes is equal to 79% and for the homologous genes is equal to 93%.

Table 4.2. Number of coding sequences in two different datasets of honeybee coding sequences.

	Number of sequences with GC < 50% (biased)	Number of sequences with GC > 50% (unbiased)	X ² Value	p Value
Honeybee Coding sequences Homologous to Rice High GC content Coding sequences (E=1E-20)	263 (93%)	19 (7%)	35.21	< 0.001
Honeybee total Coding sequences	5302 (81%)	1249 (19%)		

Table 4.2 compares the number of sequences with GC contents below or over 50% (i.e. unbiased) in both total sequence dataset and the subset of coding sequences which are homologous to the high GC content rice genes.

In other words in both species homologous sequences with on average shorter lengths are at both extreme sides of the GC content distribution histogram. The directions of these extremities are the same as the direction of the bias in the nucleotide content of the entire genomes. For example, in honeybee with a GC poor genome, this subset of genes are more leaned towards lower GC contents while the same subset of genes in rice genome are highly GC rich (as I picked them at the first place, i.e. GC contents greater than 60%). Thus apparently there has been some sort of selective constraint acting more on those longer genes and less on the shorter genes,

allowing the latter set to accumulate more changes and show more biased characteristics. I will discuss this more in discussion section.

4.4. Discussion

The debate between researchers who believe in the greater importance of mutation pressure (and then fixation of the mutations in a population through genetic drift) and those who believe in the greater impact of the selection in evolution is a very long aged debate that has been going on for decades. The theories explaining the evolution of the genomic nucleotide content are no exception either. There are two major groups of theories explaining the very heterogeneity in the genomic nucleotide content, both inter and intra species. These theories are based on either the greater importance of mutation bias or the selective constraints acting on the genomic level shifting the values of the genomic nucleotide content. However, the results of my study on honeybee genome and comparing them with the previous results on rice genome achieved in our laboratory unveiled a negative correlation between gene length and deviation from unbiased nucleotide content. These findings lead me to a new explanation on the causes behind the heterogeneity in nucleotide contents in which can be considered a neutralist-selectionist model. In my suggested model, a higher chance of negative selection against longer genes in a situation with biased mutational pressure can lead to heterogeneity in genomic nucleotide content within a genome. My model emphasises on the susceptibility of longer genes comparing to shorter ones and as a result a slower rate of substitution (i.e. accumulation and fixation of changes) in the longer genes. My model can also explain the correlation between recombination rate and GC content evolution as explained by Fullerton S.M. *et al.* (2001) and Meunier J. and Duret L. (2004). This model which is in contradiction with what Comeron *et al.* (1999) claim in their studies (i.e. longer coding sequences can be affected less efficiently by selection) is, in fact,

consistent with the background selection model proposed by Charlesworth in 1993 (Charlesworth *et al.* 1993; Charlesworth *et al.* 1995; Loewe and Charlesworth, 2007; Charlesworth, 2010). Based on this model, selection against harmful mutations at a genetic site may reduce the genetic variability in the closely linked sites. This in a sense is the same as what we see in selective sweep. In the latter model, the genetic variability of the closely linked sites of a site subject to a beneficiary mutation will as well reduce but through the fixation of the mutated site and raising its fitness.

I also studied the rate of changes in different codon positions and found out that the third codon position has the highest rate of substitutions comparing the first two positions. The fact that most of the mutations at third codon position are silent mutations, lead me to the fact that the main cause behind the changes in the genomic nucleotide content in honeybee (as it was shown the same in rice by Wang *et al.* in 2004) is bias in mutation. What I show here however, is the fact that such heterogeneity in honeybee genome, can be caused by a “constant” mutational bias through the entire genome’s length, but, accompanied with the act of natural selection acting stronger on longer genes. I will discuss my results more in detail in Chapter 5 (i.e. General Discussion).

Chapter 5. General Discussion

The variation in genomic nucleotide content has been a puzzle of interest for many researchers for decades (e.g. Bernardi and Bernardi, 1985, 1986; Bernardi, 1989, 1993, 2000, 2001, 2002; Constantini *et al.* 2006). This very simple yet important feature of genomes has been shown to have various effects on different aspects of a living organism (Bernardi, 2000). Moreover, the nonstationarity in this feature has been proven to affect the accuracy of different studies such as phylogenetic studies. In previous chapters I explained some of these effects such as the effect of variation in nucleotide content on the amino acids content of an organism (Chapter 4), its effect on the substitution rates between different nucleotides as well as between purines and Pyrimidines (e.g. transitions and transversions) (Chapter 3). I also talked about the effects these heterogeneities might have on our phylogenetic studies. In general one might say that these observed heterogeneities affect the DNA and protein molecular evolution models and hence can affect the reliability and the accuracy of those studies that use these models such as phylogenetic studies.

5.1. The ebb and flow of genomic nucleotide content

The phylogenetic relationship among different species in *Plasmodium* species has been a standing debate among scientists for years now (Dávalos and Perkins, 2008). Dávalos and Perkins point out two major reasons behind the conflicts in inferring the phylogenetic relationships among the species in this lineage; saturation and bias in nucleotide composition in the species in this lineage. Among all the species in this lineage, here, I studied the bias in the nucleotide composition of two species that are responsible for human malaria; *P. falciparum* and *P. vivax*.

Even though these two malaria parasites infecting human (*P. falciparum* and *P. vivax*) belong to the same lineage, their GC contents show very different values which contradicts with the general tendency in close species to show similar GC contents. My results show that the significant difference between the GC content values in these two species is not due to a set of sequences that are specific to one of these species and runs through the entire genome of the two species (Figures 2.1 A and B). I further showed that this trend runs through the entire genome of the two *Plasmodium* species between every single pair of orthologous genes (Figure 2.2), even in the subset of very low GC sequences in *P. vivax* (Figure 2.3).

The important question to answer is how this difference in GC content arose in the first place (i.e. millions of years ago, when the two species started to diverge). Was it *P. falciparum* that started to lose its Guanine and Cytosine content? One might guess that since the GC content of *P. vivax* is closer to an unbiased value, it is *P. falciparum* that has been losing its Guanine and Cytosine content and shifting its value towards lower amounts while *P. vivax* has retained an unbiased amount since its divergence from *P. falciparum*. I found that this was not, in fact, the case.

To understand what that happened to the GC content in either species' genomes, I needed a signal more accurate than the overall GC content. For this reason, as mentioned in the methods section of Chapter 2, I calculated the GC content in different sites of these two genomes e.g. conserved and variable sites, postulating the conserved sites as a common signal between the two species which they have inherited from their most recent common ancestor (MRCA). Thus, to calculate the amount of changes in one species' sequence (more precisely sequence feature, such as GC content), I compared it with the nucleotide content of the conserved sites. Table 2.5 shows that the conserved sites between the two species show GC content fairly close to the overall GC

content in *P. falciparum* comparing the content of Guanine and Cytosine in *P. vivax*. The absolute value of the difference between *P. falciparum*'s GC content and the conserved sites is equal to 3% while this difference between the GC content of the *P. vivax* and the conserved sites is equal to 18% which is almost six fold higher. Thus generally speaking, it seems that *P. vivax* is undergoing many more changes in GC content than *P. falciparum*, at least from the time they started to diverge from each other.

According to Kimura (1980) (also shown later by several other studies) the second codon position has the slowest substitution rate. He explains his theory using the neutral theory of molecular evolution (Kimura 1968, King and Jukes 1969, Kimura and Ohta 1974, Kimura 1980). Changes at second codon position lead to changes in the coded amino acids to the new ones with greater differences comparing the situation that the changes occur in the two other codon positions. Kimura shows that when the substitution happens at second codon position, the newly coded amino acid will be more different from the original amino acid (i.e. nonsynonymous substitutions) comparing to a substitution happens in the first or third codon position. This was later shown by Sanjuán and Bordería (2001). In their study on the evolution of proteins in HIV-1 and the secondary structure of this virus's RNA, they showed that all the substitutions at second codon positions are replacement substitutions. Thus, it seems reasonable to conclude that changes at second codon positions are more likely to be harmful (and as a result to be selected against) than if the changes happen at either first or third codon positions. Kimura (1990) shows that changes at first codon position are also more prone to be harmful than the third codon position. In general one can say that changes at third codon position are the least harmful changes. In most cases nucleotide changes at third codon position will not even result in a change in the coded amino acid at all (synonymous substitutions). This phenomenon has been

described as codon degeneracy (e.g. TTN codes for Serine amino acid. N can be any of four nucleotides i.e. A, G, C, or T). Thus the chance of accumulation (fixation) of nucleotide changes in this position (i.e. third codon position) is greater than the first and second codon positions. To put this in other words, one can say that the rate of changes (i.e. substitution rate, not to be confused with mutation rate) at third codon position is faster than the other two positions. Second codon position has the least possibility to accumulate these changes since there will be a greater chance that the changes are harmful and as a result subject to removal from the population by negative selection, while most of the changes at third codon position are synonymous substitutions that can be fixed in the population by random drift (i.e. neutral theory of molecular evolution) (Kimura and Ohta 1974). Yang (1996) mentions this maintenance of the substitutions at third codon positions as this codon position contains more information than the other two codon positions. Thus different codon positions are actually considered as different sites in phylogenetic studies and some of the current methods of inferring relationships between species, in fact, use them to partition the sequence data and deal with them differently (Ronquist and Huelsenbeck, 2003).

Looking at the data in different codon positions from the two *Plasmodium* species (Figure 2.5 A and B), one would expect to see fewer changes at second codon positions from the time of their divergence compared to the other two positions. On the other hand it would not be a surprise to see more changes at third codon position. When we bring this to the GC content context, we should be able to see more changes in the GC content at third codon positions rather than the second and first one. More precisely, comparing the GC content of the total sites (i.e. conserved and variable combined) in different codon position to the GC contents of the only conserved sites (as a relatively extant evidence of the ancestral sequence), we would expect the

changes in GC content in second position to show the lowest amount while these changes at third codon position show the highest amounts. This is the same for both species *P. falciparum* and *P. vivax* and similar to what Kimura showed in his studies on substitution rates (Kimura 1980). Looking at the data for different codon positions in each of the two species genomes (Figure 2.5 A and B), it is very clear that in both cases the differences between the average GC content in the conserved sites (Figure 2.5 B) comparing to all sites (Figure 2.5 A), is more pronounced at third codon position. The first codon position is the runner up and the lowest amount of differences can be seen at second codon position. This is correct in both species. In *P. falciparum* the changes in GC content at first codon position and second codon position are equal to 2%, while this change at third codon position is equals to 9%. In *P. vivax* these values are as 12% at first codon position, 11% at second codon position, and 32% at third codon position. My results showed to be consistent with Kimura's neutral theory of molecular evolution. This so far, can be true about all the species' genomes no matter what genomic GC content they have, because it is only a result of different rates of substitutions among different codon positions. In this study however, not only did I show this characteristic, I also reveal the fact that when I compare the mentioned variation in GC content in different codon positions between the two *Plasmodium* species to the conserved sites, *P. vivax* clearly shows higher amounts of changes in each of the codon positions than its corresponding codon position in *P. falciparum* (i.e. changes in GC content at first codon position in all sites from the conserved sites in *P. vivax* is higher than in *P. falciparum*. This trend is also correct when we look at the amount of changes at second codon position as well as at third codon position). The highest variation between the GC contents of each of the two species comparing the conserved sites, of course as we expect, can be seen at third codon position. This is consistent with the higher substitution rate at third codon position

explained earlier in this section. These data clearly show that the GC content of *P. vivax* is deviating from its ancestral values faster than the GC content of *P. falciparum*.

These differences will be even bolder when I take only the variable sites into account (Figure 2.5 C). Earlier I showed that when we take all the codon positions into account, the GC content difference between variable sites and conserved sites in *P. falciparum* is equal to 9% while this amount for *P. vivax* is equal to 44% which is almost 5 times larger. Looking at variable sites in different codon positions separately (Figure 2.5.C), we can see that the amount of changes in the GC content at first codon position for *P. vivax* is equal to 34% while this amount for *P. falciparum* is equal to 7% (almost 5 fold smaller). At second codon position, the changes in GC content in variable sites in *P. vivax* is equal to 42% while this amount in *P. falciparum* is equal to 7% (6 fold smaller). Looking at third codon position, the changes in GC content in variable sites in *P. vivax* is equal to 49% while this amount for *P. falciparum* is equal to 14% (3 fold smaller).

The lower rate ratio of changes at third codon positions between the two species comparing this rate ratio at second codon position can be caused by saturation at third codon position as this phenomenon was pointed out by Dávalos and Perkins (2008) as one of the reasons behind the existing conflict in *Plasmodium* phylogenomics. In general, one can say that the differences between GC content values in total sites in *P. vivax* and of those values in conserved sites are much higher than the differences between GC content values in total sites in *P. falciparum* and of those in conserved sites. This is true in all different codon positions and for different sites (variable or total sites) regardless to the rate of nucleotide changes at those sites and codon positions. Only in those sites with greater rates of substitutions, the difference between these values becomes more pronounced. In simpler words, in *P. vivax* the genomic content is deviating

from its ancestral state faster than *P. falciparum* and this deviation gets even faster when we look at those sites or codon positions with greater substitution rate.

These results so far clearly show that *P. vivax* is going under much more dramatic changes than *P. falciparum* in terms of its genomic GC nucleotide content. But how is it that we see a fairly unbiased amount of GC content in the today's genome of *P. vivax*? One can explain this results in a way that through the course of evolution, the early ancestors of the two *Plasmodium* species have undergone changes that have driven their nucleotide content towards very low GC contents, but after the divergence of the two *Plasmodium* species, around ten millions years ago, *P. vivax* has started gaining back its GC content in a very dramatic pace. In other words, the genome with biased directional changes in its content would be *P. vivax* rather than *P. falciparum* while *P. falciparum* is retaining its nucleotide content since the divergence from its most recent common ancestor with *P. vivax*. This is indeed important in tuning our phylogenetic analysis methods based on the bias in the nucleotide content of these species' genomes and also the direction of the bias in those genomic contents. My findings might come in handy solving the ongoing debate on the origin of the two *Plasmodium* species; *P. falciparum* and *P. vivax*.

The next question was: are the individual nucleotides (i.e. A, G, C, and T) contributes equally in the changes seen in the nucleotide content or are these changes driven mainly because of one or two specific nucleotides. The results of the calculation of each individual nucleotide content and the comparison between these contents in conserved and variable sites at different codon positions (Figures 2.6 and Appendix Figure 2.1 A and B) helped me to dig deeper into this question.

Comparing the content of each of the individual nucleotides in *P. vivax* with their corresponding values in conserved sites between *P. vivax* and *P. falciparum* shows that the changes in the GC content are not evenly distributed between the two nucleotides Guanine and Cytosine in *P. vivax*. The same is true when we look at the changes in *P. falciparum*. More precisely, in *P. vivax*, apparently Cytosine and Adenine are contributing more to the changes (i.e. increase) in the GC content than the other two nucleotides while in *P. falciparum* the changes in two nucleotides Thymine and Guanine play the major role in decreasing the GC content of this genome comparing the other two nucleotides. In other words the faster rates of changes in the number of Adenines and Cytosines in *P. vivax* suggest that, in this genome, Adenine is being replaced by Cytosine at a faster rate comparing the other possible types of substitutions leading to a higher GC genome. In *P. falciparum* however, this is true about Guanine and Thymine. In other words, in this genome, apparently the rate of Thymines replacing Guanines is faster than other possible changes leading to a lower GC genome. However, these trends do not seem to be constant among different codon positions.

The results of all these comparisons can give us more realistic models of DNA evolution in these species and hence can be very helpful for modifying the current phylogenetic analysis approaches to more accurate methods with more sensitivity towards biases in nucleotide contents. The fact that I found that *P. vivax* is gaining back its GC content at a very fast rate shows that this feature of the genome of a species (i.e. changes/bias in the genomic nucleotide content) is not always happening in one direction and can shift back and forth during evolution, probably due to some environmental changes such as a change of a habitat (e.g. host switch in parasites). This phenomenon, which I call “the evolutionary ebb and flow of genomic nucleotide content” can affect some other features of a species such as their amino acid content. Also, my

results showed that the current state of a genome's nucleotide content cannot always address the direction of the bias happening in that genome, as we saw that this was not the case in *P. vivax*. This shows that researchers need to consider the history of the changes in the nucleotide content of a genome along with the current state of this content in order to have a more accurate phylogenetic inference result.

5.2. Distortion of the usual transition-transversion ratio

Different models of molecular evolution of DNA define the rates of substitutions between different nucleotides as well as the frequency of each of the nucleotides. These models are important because of their role in correcting the methods of calculating genetic distances between nucleotide sequences that are used in phylogenetic inference studies (Yang and Rannala, 2012). In a simple model such as Jukes-Cantor model, which considers a unique rate of substitutions between any two nucleotides and a similar frequency for all four different nucleotides (Jukes and Cantor, 1969), this distance can be easily calculated using the number of non-identical nucleotides between sequences, however if the rate of changes between different pairs of nucleotides happen to be unequal, this calculation will be more complex and that is where the other models of DNA evolution come in handy.

Almost all of the different models of DNA evolution consider that substitutions between two nucleotides from the same structural class (i.e. Purine or Pyrimidine) happen with a greater rate than those between nucleotides from different classes (Wakeley J. 1996). There is one exception to this general rule, the “Jukes-Cantor” model. In Jukes-Cantor model (Jukes and Cantor, 1969) that considers all the changes to occur equally likely and there is no bias towards any specific substitution, the ratio of the rates of transitions over transversions is equal to 1:2. This is simply because there are twice as many possible transversional changes as transitions. In reality,

however, the rate of transitional changes is not, in fact, the same as transversional ones and thus the 1:2 rate ratio does not seem to be a very realistic ratio. This ratio in other models, that state the existence of transitional bias, is different from Jukes-Cantor's 1:2 ratio. In Kimura's two parameter model (K2P) (Kimura M. 1980) this ratio is defined as $\alpha:2\beta$ considering different rate for transition (i.e. α) than for transversion (β). Since there are twice as many possible transversion types as transitions, this ratio is noted as $\alpha:2\beta$. It is clear that once $\alpha=\beta$ (i.e. no bias towards either transition or transversion), this model depicts the same rate ratio as Jukes-Cantor model (i.e. 1:2).

There are a handful of other models for molecular evolution of DNA, but what was interesting to me in this research was the inequality between the rates of transitions and transversions caused by the bias in the nucleotide content. More specifically, how bias in nucleotide content can distort the usual transition/transversion (ts/tv) ratio was the puzzle of my interest.

I chose the two mitochondrial genomes of *P. falciparum* and *P. vivax* in order to study the possible effect of the bias in nucleotide content on ts/tv ratio. The reasons for choosing these two datasets were; first they show a very low GC content (i.e. biased nucleotide content). Also, as Wakeley (1996) showed, the transition bias is more pronounced in mitochondrial genomes comparing the nuclear (or chloroplast) genomes. Another reason to choose these two datasets was because of their long divergence time, I assumed that they have had enough time to show the effects of different evolutionary forces on their genome content. In the meantime, they are not very distant species and this helps me to avoid saturation effect through multiple changes at the same site. Saturation in substitution has been shown before that will decrease the phylogenetic information (Xia *et al.*, 2003). However, I tested my datasets for saturation.

Comparing these genomes, my results indeed showed a different (in fact, opposite) behaviour in terms of the bias in the ts/tv ratio. Looking at Figure 3.2 (A and B), we clearly see a very large number of transversions comparing the number of transitions, which contradicts the usual ts/tv rate ratio (Wakeley, 1996). The ratio of ts/tv between these two mitochondrial genomes is equal to 0.61:1. As it was shown in Figure 3.2, among transversional changes between these two genomes, those between Adenine and Thymine compose the largest part. This trend is simplified in a schematic cartoon (Figure 5.1). Comparing this figure with the usual trend (Figure 3.1) the difference is very clear.

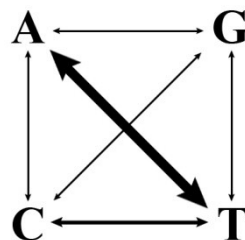


Figure 5.1. Schematic representation of nucleotide substitutions between two mitochondrial genomes of *P. falciparum* and *P. vivax*. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

It is worth noting that these two mitochondrial genomes show a very low amount of GC content values (i.e. 27% for *P. falciparum* and 25% for *P. vivax*) (Supplementary Table S3.1). From the same table, we can see a fairly low amount of GC content in the conserved sites between the two species (i.e. 30%). I conclude that the most recent common ancestor (MRCA) of the two species also used to have a low GC content. This shows that there are evolutionary forces trying to keep the GC content in these two genomes very low. These forces constraining the GC content might include mutational pressure, selection pressure, or even interplay between the two.

There are more detailed explanations about the mechanisms of these evolutionary forces in Chapter 1. On the other side, as explained in the introduction section of Chapter 3, there are evolutionary forces (both in mutational level and selection against replacement mutations) that try to keep the bias towards higher rates of transitions. As it was mentioned earlier in this chapter, based on the mitochondrial codon table and taking only third codon position into account, one can clearly see that at this codon position, a higher percentage (i.e. 43%) of the transversions are replacement (i.e. non-synonymous) mutations comparing to the transitions in which all of them are synonymous. Thus one might expect that evolutionary forces constraining the amino acid composition to limit the number of transversions. I also showed that between these two genomes, 94% of the transversional changes at third codon position are those between Adenine and Thymine. Then we can expect that those forces constraining the composition of amino acids, to extensively limit the number of changes between Adenine and Thymine. But my results show that, in fact, this is not the case. Putting these together, alongside with the GC content data in two *Plasmodium* species (Supplementary Table 3.6) leads me to the conclusion about what has been happening in these two genomes since the time of their divergence.

If we assume that since the divergence of these two species from their low GC content ancestral mitochondrial genome, the changes between their mitochondrial genome have been mostly flipping between Adenine and Thymine nucleotides it will satisfy both the maintenance of the very low GC content as well as the very high number of transversional changes between nucleotides Adenine and Thymine. This is even more significant when we take only the third codon position into account where more transversions are allowed to happen (because of less selection pressure).

Although it will not be a surprise to see a relatively higher rate ratio of transitions over transversions at both first and second codon positions comparing the third codon position. This might be because of larger selection pressure at these codon positions. If we look at the codon table we can clearly see that most of the changes at these two codon positions are replacement changes as opposed to synonymous changes.

Because of the reasons explained above, one might expect a higher rate of transversions at third codon position comparing the first two codon positions. But interestingly even in this situation, if we exclude the changes between nucleotides Adenine and Thymine, the number of transversional changes drops to as low as 13 changes at this codon position. This shows that the forces constraining the transversional changes are very strong in limiting the number of transversions even at third codon position although apparently in the case of limiting the number of A \leftrightarrow T transversions they have been overridden by other evolutionary forces (i.e. forces trying to keep the GC content low). In other words, the forces keeping the rate of transversions low have been able to affect all transversion types except the changes between Adenine and Thymine which were under the stronger effect of those evolutionary pressures keeping the GC content low. This as well happens at first codon position where the transition rate is higher than transversion rate and 63% of the transversions are changes between Adenine and Thymine, but not as dramatic as what that happens at third codon position. Although at second codon position, which is the most conserved codon position, the changes between Adenine and Thymine represent a fairly normal behaviour and compose around one fourth of the possible transversional changes (i.e. 29%). This might be because this codon position is the most susceptible to selection and any changes at this codon position leads to changes in amino acids. Thus apparently evolutionary forces constraining the amino acid composition at this codon position are stronger

than those keeping the GC content low. Although we saw that this was the opposite at third codon position. Considering the results in all codon positions combined, one can conclude that in this case (i.e. mitochondrial genomes of *P. falciparum* and *P. vivax*) the resultant force in all codon positions is in favour of those forces keeping their GC content at a low amount after their divergence. In other words, after the divergence of these two species, the magnitude of the forces keeping them low GC has been overriding selection pressure constraining the amino acid composition (which in turn keeps the bias towards transitions). Thus, basically what I show here is that when the bias in the genomic nucleotide content in two genomes happen to be towards the same direction, the bias will distort the usual *ts/tv* rate ratio because of flipping between two nucleotides from different classes of the nucleotides (in this case between Adenine and Thymine). It will be interesting to check out two species with bias in their GC contents towards high values to see if we can see the same distortion happening and if it does, is it because of the flipping between Guanine and Cytosine?

Looking at the substitution matrix at different codon positions between these two mitochondrial genomes, one can notice that the most part of the “flipping” between Adenine and Thymine has been happening at third codon position (Supplementary.). This shows that the mutational pressure is mostly responsible in keeping the GC content at a low value and as a result to change the *ts/tv* rate ratio. Studying the *ts/tv* rate ratio behaviour is important in phylogenetic studies because it is considered as one of the parameters in the DNA molecular evolution models along with the frequency of the four different nucleotides in many phylogenetic inference methods such as Markov Chain Monte Carlo (MCMC) (Yang and Rannala, 2012).

Substitution saturation is one of the problems to deal with in phylogenetic studies (Dávalos and Perkins, 2008). It was pointed out by Perler (1980) studying the evolution of preproinsulin and globin in chicken. She realized that the accumulation of synonymous changes is not linear as it is for replacement changes. This phenomenon, which was studied broadly by Xia and his colleagues (Xia *et al.* 2003; Xia and Lemey, 2009) happens when the rate of the changes is high enough or the divergence time is long enough so one single site gets the chance of changing more than one time. Sometimes these further changes even can change a nucleotide back to its original type (e.g. A→T→A). Thus it is clear that if a sequence gets to a saturation point, it will start losing its phylogenetic information value (Xia *et al.*, 2009). While sequences reach the full saturation point, the similarity between them will be only affected by the frequency of the different nucleotides and we know that this cannot be a very accurate phylogenetic signal to count on (Xia *et al.*, 2003). The higher the rate of changes in one sequence/site is the faster it can reach the saturation point. Thus one might expect that the saturation happens faster in sites subject to transitions than those with transversional changes. Also, for the same reason, among different codon positions, it will be reasonable to expect faster saturation at third codon position compared to the first two codon positions. However, my focus in this study was on the sites subject to transitional changes and will compare them to those with transversional changes.

As I explained earlier in the introduction of Chapter 3, mitochondrial genomes are shown to have different characteristics than nuclear genomes in terms of their faster evolution (5 to 10 times) than nuclear genome (Brown *et al.*, 1979; Brown *et al.* 1982; Ingman *et al.*, 2000; Ballard and Whitlock, 2004). I also mentioned the great transition abundance over transversions in these genomes (Brown *et al.* 1979, Brown *et al.* 1982). However, this story seems to be different in plant mitochondrial genomes (Galtier, 2011). Brown and his colleagues also showed that there is

a chance of saturation when two species are phylogenetically farther from each other. Thus to study the possibility of saturation happening in mitochondrial genomes of *P. falciparum* and *P. vivax* since their divergence time, I compared the *P. falciparum* mitochondrial genome with a very closely related species mitochondrial genome. The reason for choosing a close species to compare with *P. falciparum* is since they haven't had enough time since their divergence, multiple changes at one site between the two genomes will not be very likely to have happened. Thus comparing the results from this comparison with the one comparing *P. falciparum* and *P. vivax* will give me an idea on saturation happening in the latter case. I chose *P. reichenowi* because it is the closest species to *P. falciparum* and also because it shows a relatively low GC content in its mitochondrial genome (same as *P. falciparum*). The *ts/tv* ratio between *P. falciparum* and *P. reichenowi* is equal to 1.24:1 (61 transitions as opposed to 49 transversions). Among the transitions, the changes between Cytosine and Thymine compose the majority of the changes (44 cases). This trend is shown in a schematic cartoon in Figure 5.2.

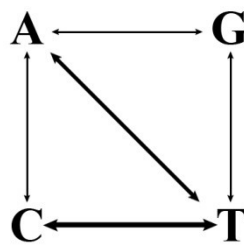


Figure 5.2. Schematic representation of the nucleotide substitutions between mitochondrial genome of *P. falciparum* and *P. reichenowi*. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

Figure 5.3 shows a schematic graph of how the substitution saturation happens. Assuming two sequences with two different substitution rates, one 10 times faster than the other one (in this graph, the thick dashed line represents the sequence with 10% substitution rate per site per million years while the dotted line represents another sequence with 1% substitution per site per million years). I use the percentage of sequence difference in this schematic figure since it is the one feature that can be easily calculated comparing two sequences (e.g. a sequence before and after changes). As it is clearly shown in this figure, at the beginning, the percentage sequence difference in the sequence with faster substitution rate takes off very fast (reflecting its substitution rate) but after a while (in less than 20 million years) it starts to show a slowdown in its increase rate (and diverging/underestimating the substitution rate at these sites). Meanwhile however, the other sequence with slower substitution rate shows a rather slow but steady increase in the percentage of sequence difference through the entire 100 million years shown in this figure. As it is clear in this figure, even though the rate ratio will stay constant through the time and no matter how far in the time scale we look into it, the percentage sequence difference values in these two sequences eventually will converge some time along their evolution.

The percentage sequence difference at the time they converge will be somewhere less than 75%. This 75% can be easily calculated if we use Jukes' and Cantor's formula for calculating the actual number of substitutions per site based on the observed number of substitutions per site (i.e. percentage sequence difference) (Jukes T.H. and Cantor. C.R. 1969). Based on their formula, the percentage of the observed substitutions will be always less than 75%. Thus no matter how many substitutions happen in different sites of each sequence the percentage sequence difference between two sequences will not reach 75% or higher. This is due to the possibility of multiple changes happening at the same site.

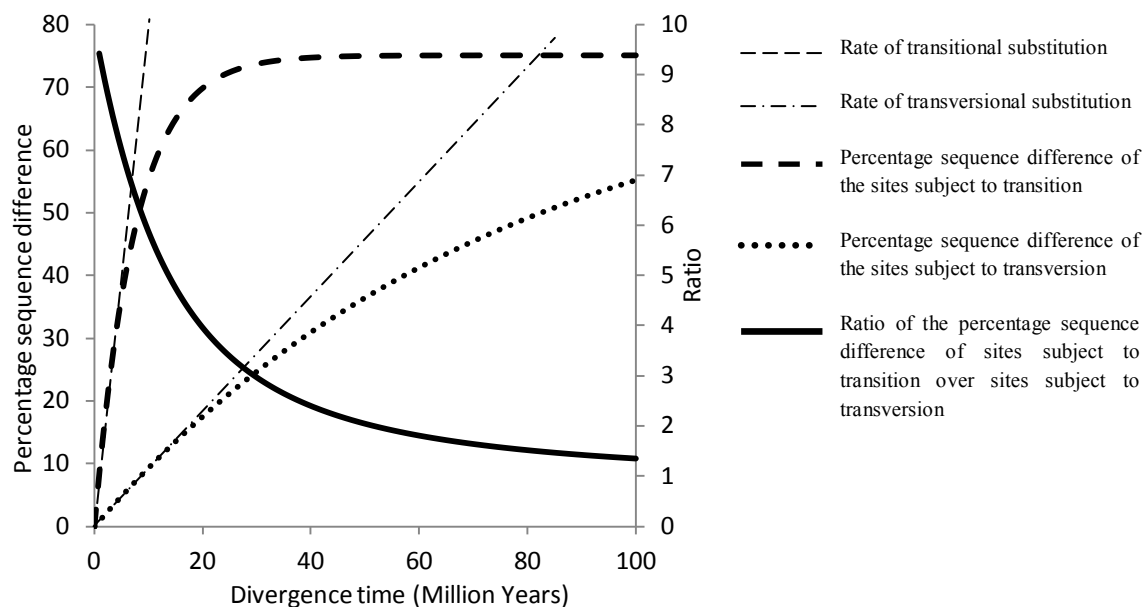


Figure 5.3. The relationship between the percentage sequence differences and the divergence time based on Jukes-Cantor model in slow (dotted line) and fast (dashed line) evolving sites.

The thin dashed line represents the rate of substitution for the sites subject to transitions (fast changes) and the thin dash-dot line represents the rate of substitution in the slow changing (sites subject to transversion).

The continuous bold line in this figure shows the ratio of the percentage sequence difference of fast evolving over slow evolving sites (the second Y axis, on the right). In the case of slow evolving sites the rate of substitution is 1% per million years and in the case of fast evolving sites, this rate is 10% substitution per million years. This model can be applied to the rate ratio of transitions over transversions. It is clear that through time, despite the fact that the rate ratio of the changes in the sites subject to transitions over those subject to transversion is constant and greater than one, the ratio of percentage sequence difference (which can be calculated) between these two set of sites is constantly decreasing (the continues bold line). This happens after the two sequences start with a very low percentage sequence difference (but very high percentage

sequence difference ratio which at the beginning reflects their substitution rate ratio) but then very soon after a short while (~ 20 MY in my example) the percentage sequence difference of those sites subject to transitions reaches the maximum distance. After this point, the percentage sequence difference in sites subject to transitional substitutions reaches the saturation while in the other sites (those subject to transversions) the percentage sequence differences is still increasing until these two converge at one point (75%). This is when the ratio of the percentage sequence differences between these two set of sites is equal to one. The constant decrease in the percentage sequence difference ratio is caused by saturation happening in those sites subject to transitions (higher rate substitutions) raised from the possibility of multiple changes in one site and the fact that some of those sites might reverse the former changes and change back the nucleotide to its original type. This saturation causes underestimating the substitution rates through the decrease in the ratio between percentage sequence differences in two sets of sites.

Comparing the number of changes between *P. falciparum* and *P. vivax* with the number of changes between *P. falciparum* and *P. reichenowi* shows that the number of transitional changes between *P. falciparum* and *P. vivax* is almost three times more than the transitional changes between *P. falciparum* and *P. reichenowi*. In the mean time however, the number of transversional changes between *P. falciparum* and *P. vivax* is over five times more than the transversional changes between *P. falciparum* and *P. reichenowi*. One might think that the ratio of percentage sequence difference in those sites subject to transition over those subject to transversion has decreased to almost half between *P. falciparum* and *P. vivax* comparing the other pair of species and this can be a sign of saturation in the transitional substitutions between *P. falciparum* and *P. vivax*. But if we take a closer look at the number of changes between Adenine and Thymine, we will see the number of changes between these two nucleotides is more

than five times greater than the rest of the transversional changes combined between these two species. If we exclude this outstanding number of changes between Adenine and Thymine, forced by the nucleotide content bias, the percentage sequence difference ratio between *P. falciparum* and *P. vivax* will increase to 4.09 which is more than twice greater than this ratio between *P. falciparum* and *P. reichenowi*. This shows that however the comparison of the percentage sequence difference ratios between two cases of *P. falciparum* vs. *P. vivax* and *P. falciparum* vs. *P. reichenowi* shows that substitution saturation seems to be inevitable, but this difference is only due to an extraordinary increase in the number of changes between Adenine and Thymine between *P. falciparum* and *P. vivax*. As I explained earlier, this in turn, is affected by the very low GC content of the two species mitochondrial genome and flipping between Adenine and Thymine nucleotides between these two species. Thus it seems to be a naive conclusion to claim that the substitutions between *P. falciparum* and *P. vivax* has reached the saturation point just based on the ratios of the percentage sequence difference. Blouin *et al.* (1998) mention the substitution bias and accelerated mutation rate as causes of extremely rapid saturation in silent sites. I explained in Chapter 1 (Section 1.5) that in mitochondrial genomes of mammalian and invertebrates, transitional changes at third codon position are all synonymous (silent) while 43% of transversion are replacement substitutions at this codon position. Thus the excess of transversions at this codon position between *P. falciparum* and *P. vivax* mitochondrial genomes can be considered as lack of selection constraints (at least at the amino acid level). This leaves mutational bias (towards Adenine and Thymine since these two genomes are extremely AT rich) as the potentially major cause behind the extreme bias in their nucleotide content towards higher amounts of Adenine and Thymine.

Brown and his colleagues showed that mammalian mitochondrial genomes show a very strong bias towards transitions (W. M. Brown 1982, Wakeley 1996). My results also confirmed this as I found a 16.1:1 *ts/tv* ratio in my dataset. If we take a closer look at the calculated *ts/tv* ratio between human and chimpanzee mitochondrial set of genes, we clearly see that the observed transition bias is not evenly distributed between A↔G and C↔T transitions (Figure 3.4). Figure 5.4 shows a schematic representation of this unevenness.

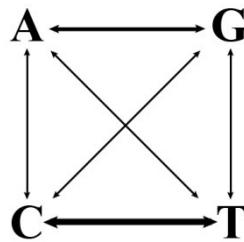


Figure 5.4. Schematic representation of nucleotide substitutions between two sets of concatenated, aligned orthologous mitochondrial genes in human and chimpanzee. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

Looking at Figure 3.4 (B), the number of changes between nucleotides Cytosine and Thymine is almost 2.5 times larger than the number of changes between Adenine and Guanine. Also (similar to what we see between *P. falciparum* and *P. reichenowi*), the number of changes between Cytosine and Thymine is almost evenly divided in both directions (117 C→T changes from human to chimpanzee as opposed to 100 C→T changes from chimpanzee to human). In other words the number of changes from Cytosine to Thymine in both species is almost the

same. One might conclude that this could be due to de-amination of Cytosines in both species which happens in a high frequency.

The bias in mitochondrial genomes in two species *P. falciparum* and *P. vivax* were towards one direction (they both are AT rich). In order to see if the bias in opposite direction can have the same effect I studied the nuclear genome of the same pair of species and calculated the ratio of transitions rate over transversions. Figure 5.5 shows a schematic cartoon of the observed rate ratio between different types of transitions and transversions.

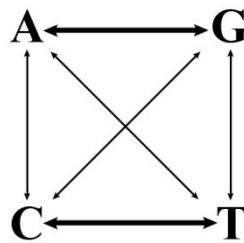


Figure 5.5. Schematic representation of nucleotide substitutions between two sets of six concatenated, aligned orthologous nuclear genes in *P. falciparum* and *P. vivax*. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

As one can clearly notice, these two subsets of nuclear genomes between two *Plasmodium* species exhibit the normal abundance of transitions over transversions. Moreover, the total number of transitions is apparently evenly divided between two types of transitions (i.e. $A \leftrightarrow G$ and $C \leftrightarrow T$).

The results of the comparison between human and chimpanzee's mitochondrial genome is similar to when I compared *P. falciparum* and *P. vivax* nuclear genomes in a sense that neither of

the cases showed a distorted *ts/tv* ratio. However, in the case of human and chimpanzee, the two datasets of mitochondrial genes show a relatively similar nucleotide content around the unbiased nucleotide content but in *P. falciparum* and *P. vivax* nuclear genome case, there were biases in both genomes nucleotide content but in opposite directions. In the latter case I witnessed unevenness in the distribution of the changes from one nucleotide to the other and vice versa arising from the bias in opposite directions. Thus one can be certain that the main cause behind the distortion of the *ts/tv* rate ratio between the mitochondrial genomes of *P. falciparum* and *P. vivax* is the “unidirectional” bias in their nucleotide content.

Comparing the nucleotide substitution rate in nuclear genome of vertebrates, we can see a relatively similar behaviour to what we saw in the nuclear genomes of *P. falciparum* and *P. vivax*. As it is schematically presented in Figure 5.6, in this case that there is no significant difference between the nucleotide content of the two fairly unbiased datasets (human and chimpanzee nuclear genome datasets), the bias is towards transitions and it is evenly distributed between $A \leftrightarrow G$ and $C \leftrightarrow T$ substitutions. This somehow resembles the usual *ts/tv* rate (Figure 3.1).

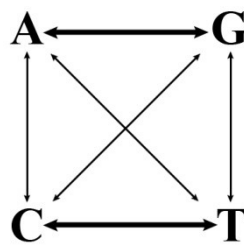


Figure 5.6. Schematic representation of nucleotide substitutions between two sets of concatenated, aligned orthologous nuclear genes in human and chimpanzee. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

So far I showed that, regardless to the far greater difference between the nucleotide contents between two *Plasmodium* species than between human and chimpanzee nuclear genomes, in both cases we still see the abundance of transitions over transversions. Also we see that in both cases the number of transitions is almost evenly divided between two types of transitions (i.e. $A \leftrightarrow G$ and $C \leftrightarrow T$). Thus one might wonder where the bias in the nucleotide contents of *P. falciparum* and *P. vivax* will show its effect on the substitution rates. If we look closer to the transitions between *P. falciparum* and *P. vivax* nuclear datasets (Figure 3.5), we can see that the “direction” of the transitional changes shows a strong bias. The direction of the observed bias is in a way that one can reasonably conclude that it is affected by the difference between the nucleotide content between the two species. This “directional” distribution of the substitutions is represented in Figure 3.7. As it is clearly shown in this figure, the number of changes from Adenine to Guanine (from *P. falciparum* to *P. vivax*) dominates the number of changes on the opposite direction. This clearly is consistent with the higher GC content in *P. vivax* nuclear genome comparing *P. falciparum* genome. The same story applies when we take the number of changes between Thymine and Cytosine into account. The number of changes from Thymine to Cytosine (from *P. falciparum* to *P. vivax*) greatly outnumbers of those on the opposite direction.

To summarize, I can conclude that in the case of nuclear genes between *P. falciparum* and *P. vivax*, the bias in the nucleotide content which is in opposite directions between the two species is still affecting the ratio of substitutions, but this effect is happening at a different level. It only changes the ratio of the changes between opposite directions of the same substitution type but keeps the ts/tv ratio as a total to its usual value. Figure 5.7 depicts this idea in a schematic cartoon.

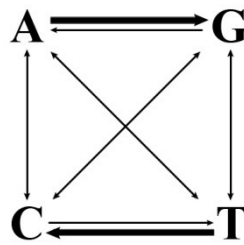


Figure 5.7. Schematic representation of nucleotide substitutions between two nucleotides from the same class, between two sets of concatenated aligned orthologous nuclear genes in *P. falciparum* and *P. vivax*. The thickness of each arrow depicts its relative abundance comparing other types of substitutions.

Here as well, $A \rightarrow B$ represents changes from any nucleotide (A) from *P. falciparum* nuclear genome to another nucleotide (B) in *P. vivax* nuclear genome (same as Figure 7).

Among those cases I studied, the comparison between human and chimpanzee nuclear genome seems to be the least biased case showing the usual ts/tv ratio (higher rate of transitions comparing transversions). Also as I showed in the results section of Chapter 3 that the number of transitions and transversions are equally divided between different types of changes, as well as between opposite directions. In a nutshell, these genomes show a very typical behaviour in terms of substitution rates. I also explained earlier that Kimura's two parameter model considers same frequencies for all four different types of nucleotides but different rates of transitions and transversions. This looks very similar to what I described here about nuclear genomes of human and chimpanzee (different rates of transitions than transversions and fairly unbiased nucleotide content in both genomes). Thus one might expect that if we compare the results from these two

genomes with the values calculated based on Kimura's two parameter model, the results should not be very far from each other. To calculate the expected values for each type of substitution in Kimura's two parameter model instead of dividing the total number of substitutions between all six types (two transitions and four transversions), I divided the total number of transitions between two types of transitions and the total number of transversions between four types of transversions. Figure 5.8.A represents the results of this comparison. As you can see, the actual numbers of each type of substitution between these two genomes are, in fact, very close to the results of the calculation based on the Kimura's two parameter model. Thus I can say that this model of DNA evolution (i.e. K2P) can work properly to calculate the distance between nuclear sequences from human and chimpanzee. But is this also the case when we take other species genomes (such as *P. falciparum* and *P. vivax* mitochondrial genome) into account?

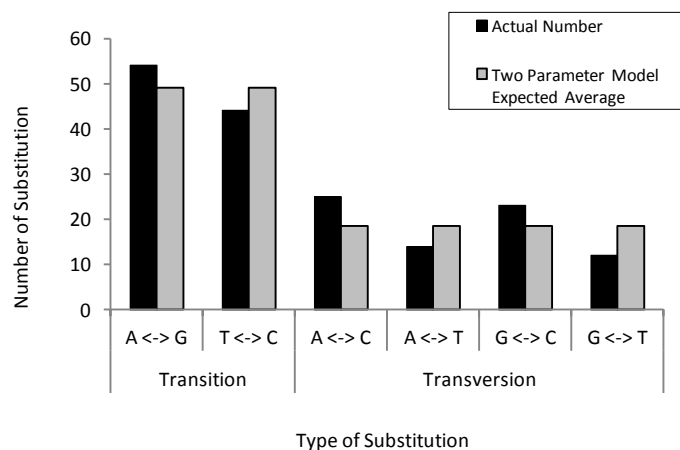


Figure 5.8.A. Nucleotide substitution pattern between nuclear genomes of human and chimpanzee. The grey bars indicate the expected average values in Kimura's Two Parameter Model of DNA evolution.

Comparing *P. falciparum* and *P. vivax* mitochondrial genomes showed the oddest results one could expect in terms of substitution rates. Moreover, the nucleotide content of these two genomes, showed a very biased (low GC) content. To challenge the properness of Kimura's two parameter model of DNA evolution to calculate the distance between mitochondrial sequences from *P. falciparum* and *P. vivax*, I did the same calculation for expected values of substitutions based on Kimura's two parameter model and compared the results with the actual observed values. The results are shown in Figure 5.8.B. It is very clear that the calculated results are far away from the actual values. This clearly shows that this model is not a proper model to calculate the distance between mitochondrial sequences from *P. falciparum* and *P. vivax* and other models (those considering different frequencies for different nucleotides) will be needed in this regard.

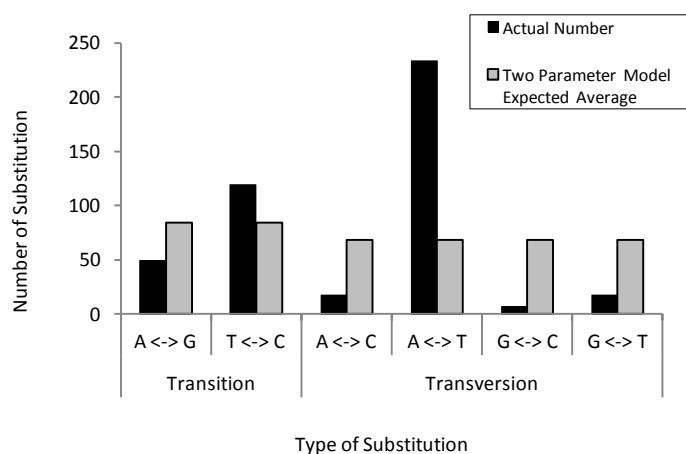


Figure 5.8.B. Nucleotide substitution pattern between two mitochondrial genomes of *P. falciparum* and *P. vivax*. The grey bars indicate the expected average values in Kimura's Two Parameter Model of DNA evolution

In Figure 5.8 (A and B) the black bars represent the actual number of each type of substitution. The grey bars indicate the expected average numbers of each substitution considering Kimura's

Two Parameter model of DNA evolution (i.e. K2P in which considers the same frequency for all types of nucleotides but different rates of transitions and transversions).

To wrap up the results of this part of my research, I can say, based on my findings, if bias in nucleotide content happens to be on the same direction between two species genomes (e.g. *P. falciparum* vs. *P. vivax* mitochondrial genomes), it will distort the usual *ts/tv* rate ratio. If the bias happens to be on two opposite directions between two genomes (e.g. *P. falciparum* vs. *P. vivax* nuclear genomes), it will not distort the usual *ts/tv* rate ratio but it will cause unevenness in the distribution of changes in opposite direction between two nucleotides (e.g. in $A \leftrightarrow G$ we will see: $A \rightarrow G \neq G \rightarrow A$ and in $T \leftrightarrow C$ we will see: $T \rightarrow C \neq C \rightarrow T$). If there is no bias whatsoever in either same direction or opposite direction (e.g. human and chimpanzee nuclear genomes), we might expect the usual *ts/tv* rate ratio. Although in human vs. chimpanzee case, when we take their mitochondrial genomes into account, the *ts/tv* rate ratio will be much bolder than when we look at their nuclear genome which is consistent with the higher rate of substitutions in animal mitochondrial genomes (Brown *et al.* 1971; Ingman *et al.* 2000; Ballard and Whitlock, 2004) and also the great transition abundance over transversions as stated by Brown and his colleagues in 1979 and 1982.

5.3. The relationship between gene length and nucleotide content

Wang *et al.* (2004) suggested the mutation bias as the main cause behind the high GC content of rice genome. They based their theory on the higher rate of synonymous changes between the two orthologous subset of genes between rice and *Arabidopsis thaliana* comparing the nonsynonymous changes (as evidence of selection acting at the protein level). Wang and Roosnick in 2006 suggested that the abundance of Guanine and Cytosine nucleotides at third codon position in Poaceae (e.g. rice) can be partly explained by selection on codon usage.

However, the negative relationship between average gene length and GC content in rice genome found by Wang and his colleagues in 2004 (same study explained earlier in the same paragraph) raised three questions (explained in Chapter 1, section 1.6) for me in which could lead me to an idea about the mechanism behind the heterogeneity in genomics nucleotide contents both between different genomes as well as within a single genome. The answers to these questions could also clarify the cause behind the observed tendency in shorter genes in rice genome to have high GC contents. Briefly these questions were: 1- If this negative relationship is a property of genomes in general or is it specific to monocot plants. 2- If the observed negative relationship is between gene length with GC content *per se* or what matters is, in fact, the degree of nucleotide bias? 3- If the evolutionary forces causing the heterogeneities, act on short genes to make them GC richer or if they act on longer genes to keeping them from becoming GC rich.

In order to answer these questions, I chose honeybee as my model organism as explained in the introduction section of Chapter 4. Even though honeybee genome is a GC poor genome (mean = 39.8%), the distribution of the GC content of the genes in this species resembles rice genome in the sense that it shows a bimodal distribution (continuous distribution with two distinct peaks) (Figure 4.1). However Serres-Giardi *et al.* (2012) claims that bimodal distribution of the genomic nucleotide content only happens in GC rich genomes (such as rice and other plants they studied) but my analysis showed two very distinctive subsets of genes in AT rich genome of honeybee with significantly different GC contents. This wide variation and bimodality (showing two distinct peaks in the distribution) in GC content in honeybee genome has been shown by other groups before (Jørgensen F.G. *et al.* 2007). This bimodality of the GC content distribution, gives me the chance to divide its entire set of genes into two subsets of genes with high and low GC contents and compare them in order to reveal the mechanism behind

this variation. Wang and his colleagues studied this feature in rice in 2004 and showed that there is a negative relationship between GC content and the average length of the genes (Wang et al. 2004). In my study I found that this negative relationship exists between the average length and AT content of the genes. Honeybee has an AT rich genome and this means that the mutation bias in this genomes is towards AT nucleotides whereas this bias in GC rich genome of rice is towards GC nucleotides. Comparing the negative relationship between the average length of the coding sequences and the nucleotide contents in these two species, one can say in both species there is a negative relationship between the average length of coding sequences and the bias in nucleotide content.

Comparing the distribution of GC content in a subset of coding sequences in honeybee which are orthologous to high GC genes in rice genome, with the total set of coding sequences in honeybee, shows that even though these sequences are orthologous to high GC content sequences in rice, they still are missing a part of genes with high GC contents. In other words, this subset of genes in honeybee is more leaned towards lower GC contents comparing the total set of genes. In other words I can say that this set of orthologous genes are showing two different extreme values of GC content in two different species. Moreover I showed (Table 4.1) that biased sequences in both species are on average shorter in length than the unbiased sequences. Some other researchers also showed the correlation between gene length and GC content in various species. Duret and Mouchiroud (1999) studied the correlation between gene length and codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis* and showed that in these genomes the higher the GC3 (GC content at third codon position gets the shorter the genes will become. Stoletzki (2011) also showed the negative correlation between gene length and optimal codon usage in yeast. Serres Giardi and her colleagues also showed (2012) this negative correlation

between gene length and GC content (specifically GC3) in 69 species (including 10 *Poaceae* species) out of 78 species they studied. This, alongside with the fact that, in opposite to what we see in rice genome, in honeybee genome there is a mutational bias towards making more ATs (the magnitude of the changes is greater at third codon position as it is shown in Figure 4.4), shows that no matter what direction the bias in nucleotide content is pointing to, once we have more biased mutations, genes are on average shorter in length (Figure 5.9).

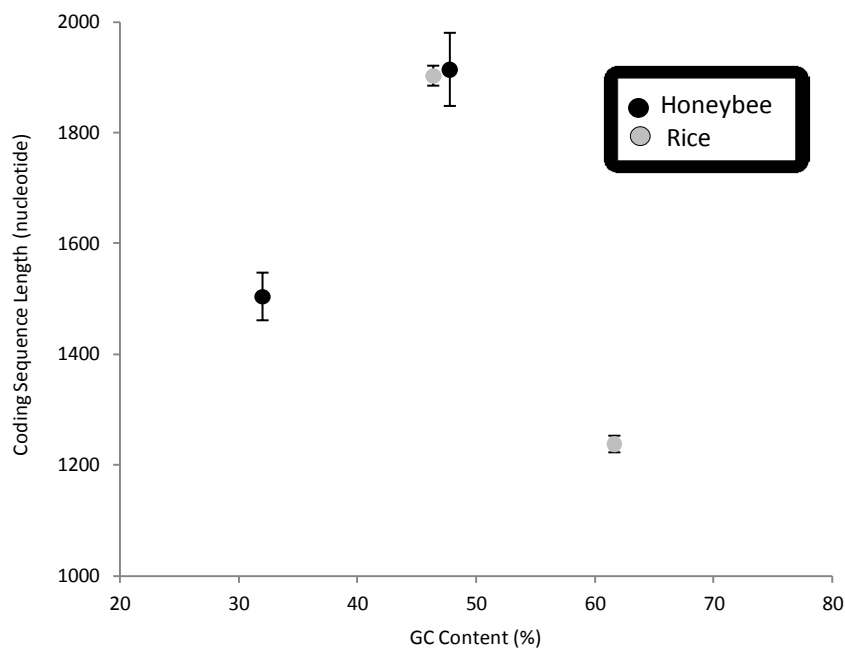


Figure 5.9. Average length of coding sequences in two groups of genes in honeybee and rice.

Sequences are sorted based on their GC content and then divided into two groups of same number of sequences, one with lower GC contents and those sequences with higher GC contents (explained in Methods Section). The confidence intervals are calculated for alpha (significance level) = 0.05 (i.e. the confidence level is 95%). The darker dots represent honeybee genes and the lighter ones represent rice genes.

To put this in other words, I can say that during evolution, affected by biased mutation, shorter genes move faster in the GC spectrum than the longer ones (i.e. show more biased GC contents). Figure 5.10 simplifies what that have been said here.

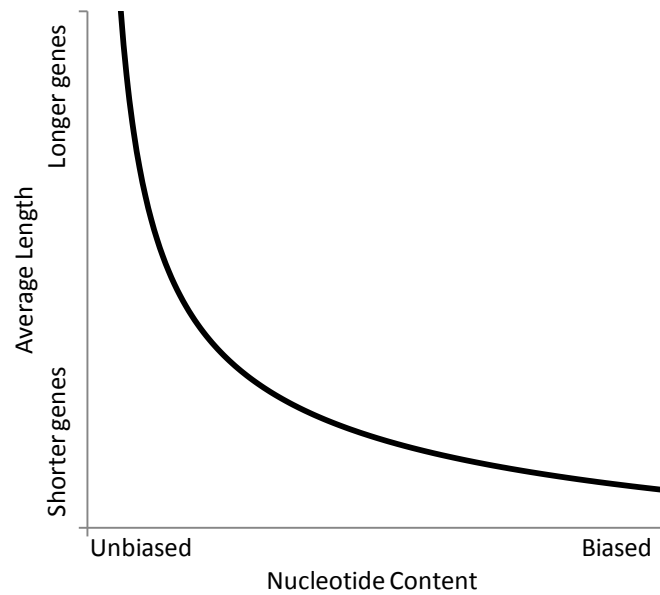


Figure 5.10. The negative relationship between bias in nucleotide content and average length in genes.

In this figure I used the inclusive term nucleotide content bias instead of high GC content or low GC content *per se* to show the general trend between the bias in nucleotide content and the average length of the genes.

I also showed that this feature is not exclusive to coding sequences and can be seen in introns as well (Figure 4.3 B and C). The more dramatic changes in introns average length can be due to their longer length and the fact that in general they occupy a larger part of each gene than exons.

The negative relationship found between nucleotide content bias and the average gene length in honeybee data alongside with rice data in which my study confirmed what Wang *et al.* had found (2004) suggests that during evolution, the accumulation of changes (fixation) in longer genes have been slowed down. In other words changes happening in longer genes haven't had the chance to be fixed in the population. Simply put they have been more susceptible to selection, and consequently, "elimination" from the population.

Let us assume we have two genes with different lengths both in the same situation in terms of the spectrum of mutations hitting them. If these mutations are biased towards more making AT (e.g. Cytosine deamination feedback loop), one might imagine that the longer gene can accumulate more changes (GC→AT) since this sequence contains more sites in which can potentially mutate to AT, and consequently can shift its GC content faster comparing the smaller genes. However, considering the fact that a part of the mutations are harmful mutations, the same long gene also can provide a larger and more accessible target for the harmful mutations. In other words the odds of getting eliminated from the population by selection for longer genes are higher since they provide larger targets for harmful mutations. My results are contradictory with what Comeron *et al.* showed in their study (1999). They claim that synonymous substitution rates are higher in longer coding sequences. They also claim that these longer coding regions have a lower codon bias. They suggest that longer coding sequences can be affected less efficiently by selection. My results however, show otherwise. In both rice and honeybee we see shorter genes to be less affected by selection and to have more chance of fixation of mutations. Once a site is hit by a harmful mutation, the individual will be removed from the population and along with the individual, all the other sites (regardless of the fact that they have not been hit by harmful mutations) will be removed from the population as well if they are close enough to the mutated

site that the recombination rate is very small in that region. Basically here I claim that the reason these “neighbouring” sites are eliminated as well is because they were “on the same boat” with the mutated site which was hit by harmful mutation. Moreover, since they were close enough to the mutated site, there was a very small chance of shuffling of the sites through recombination. However there will be two issues to pinpoint here. First the fact that in addition to harmful mutations, a fraction of mutations can also be beneficial to the individual and increase its fitness and the chance of survival and producing offspring. Thus one might say that, being a large target is not always bad and it actually might increase the chance of survival of the linked sites in the population. This is true but as it was explained in Chapter 1.5, the rate of beneficial mutations is very low comparing the harmful ones (Sniegowski et al. 2000).

The second point is the fact that when a site is hit by a harmful mutation it only can affect its linked sites (sites on the same boat) as long as nothing shuffles their line up. In other words if the linkage between two sites breaks (e.g. because of recombination), the site in which was hit by a harmful mutation cannot affect the other sites in which are not linked to it anymore. Thus based on my hypothesis one could expect that recombination rate also should affect the behaviour of the changes in nucleotide content in a way that higher rates of recombination should cancel out the effect of gene length on bias in GC content. In fact there are some studies confirming the effect of recombination rate on the evolution of GC content within a genome (Fullerton *et al.* 2001; Duret and Arndt, 2008). Moreover, Nabholz *et al.* (2011) and Birdsell (2002) studied the correlation between the GC content and recombination rate in birds and yeast and found a positive correlation between the recombination rate and GC content in these species. Serres-Giardi and her colleagues (2012) also found the same significant positive correlation in rice, maize, and *B. distachyon*. In these studies they suggest either recombination acts directly via

GC-biased Gene conversion (affecting biased mutation rate as explained in Chapter 1) or indirectly via affecting the efficiency of selection (Charlesworth 1994). Changes in GC content caused by natural selection can be cancelled out by random genetic drift unless high recombination rate overcome the effects of random genetic drift (“Hill-Robertson effect”; Hill and Robertson 1966). Thus in this approach the correlation is determined more based on selection rather than mutation (Bernardi and Bernardi, 1986). Fullerton S.M. *et al.* (2001) lists the two mentioned suggested models but is not able to explain which one of the two models are correct. In fact they confirm the existence of the correlation between GC content and the recombination rate in a genomic scale at DNA molecule level but suggest that there might be a third “unknown” factor such as chromatin structure relating these two features. Meunier and Duret (2004), along with some other researchers have also found strong correlation between GC content evolution and the recombination rate. Meunier suggests the GC biased Gene Conversion as the probable cause of the correlation between the evolution of GC content and recombination rate. Here in this study I propose a third scenario of how recombination rate can affect the nucleotide content of a genome.

It has been shown that if two loci are closely linked to each other, effectiveness of selection at one of them could also have an effect at the other locus (Nurminsky 1998). This effect can be stronger if the recombination rate between the two loci is small enough and can be reduced if the recombination rate is big enough (Felsenstein 1974). This effect can be seen in two different directions. Selection of a favoured mutation that leads to fixation of that certain mutation and can affect the genetic variation in the surrounding area, even though other loci around that mutation are neutral. This is called “Genetic Hitchhiking” or “Selective Sweep” (Smith and Haigh 1974). On the other hand, elimination of a deleterious mutation also can affect the genetic variation of

the surrounding area. The latter phenomenon is called “Background Selection”. Charlesworth defines background selection as: “the reduction in genetic variation that occurs in the genomic area surrounding a gene that repeatedly mutates to a deleterious version” (Charlesworth *et al.* 1993; Charlesworth *et al.* 1995; Charlesworth, 2010).

In the past, the theory of background selection has been applied to chromosomal segments containing several different genes. Here in my study, I essentially apply the same theory in genome scale at the level of linked nucleotide sites within a single gene. Assume a genome undergoing biased mutations towards GC, pushing the genome to have more Guanine and Cytosine nucleotides along the entire genome. If we consider a constant rate of deleterious mutations per nucleotide site, rather than per gene, we can expect that mutated copies of longer genes will be eliminated from the population with a higher probability than shorter genes. Because of background selection, even though the harmful mutation has only occurred at one nucleotide site, the linked neutral mutations (driving the GC content) in surrounding nucleotide sites within the same gene will be also eliminated from the population. Therefore, although all the genes along the genome undergo GC making mutations, the shorter genes (in which survive longer than longer genes) respond to the biased mutational process much more quickly than the longer genes.

The fact that both exon and intron sequences show the negative correlation with the degree of nucleotide bias indicates that the deleterious mutations also affect intron sequences. Although introns play little, if any, role in gene function, it is essential to be properly spliced from the primary transcript. Thus many random mutations that would interfere with the efficient splicing of these sequences could be deleterious at the level of fitness of the organism. This explains why genes with long intron sequences respond more slowly to the biased mutational pressure.

Interestingly, in broader study Liao *et al.* (2006) found a negative correlation between the rate of evolution and intron length in mammalian protein-coding genes. They found this relationship surprising but they were unable to explain the mechanism behind it. I believe that my background selection model can provide the missing explanation.

5.4. Future Research Directions

My findings in this research can help researchers to have a broader and more accurate understanding of the genomic nucleotide content behaviour. This can help researchers to fine tune their phylogenetic inference methods to be able to solve some complicated phylogenetic cases such as *Plasmodium* lineage. My findings show that not only the current state of the nucleotide content of a genome matters in an accurate phylogenetic inference study, but its behaviour through evolutionary time is also important and needs to be considered because it can change back and forth many times and, in fact, this is what that has been happening otherwise now we must have had different species with a very biased GC content values, all at one end of the GC spectrum. My results also show that the variation in nucleotide content within a genome can be a result of interplay between biased mutation and selection. My model does not reject other models proposed so far, but it can indeed come in handy explaining some of the observed variations (especially within genomes) in which other existing models haven't been able to explain.

I can list a few suggestions for further researches in this area. First, current methods take bias in nucleotide contents into account by including the frequency of different nucleotides (and in some cases in different sites) in their models for molecular evolution of DNA. My suggestions would be based on what I show in this study. Based on my results, one can develop a new model for molecular evolution of DNA that includes not only the current state of the bias in the

nucleotide content, but also the evolutionary history of the bias. In other words, the stationarity of the DNA sequences and the direction of the bias should be considered in the new model for molecular evolution of DNA.

Secondly, my results showed that the bias in nucleotide content can also affect the phylogenetic inference studies in a different way than only through the frequency of different nucleotides. The bias can also affect these studies indirectly and thorough affecting the substitution rates between different nucleotides. Another suggestion based on these results would be to fine tune different methods of phylogenetic inference considering this feature of the genomes with a bigger importance. Thus the first part of my suggestions includes developing a model for molecular evolution of DNA that takes the bias in the nucleotide content and its direction into account, and also gives a bigger weight to the frequency of the nucleotides. Once the model is developed, we can use same methods for phylogenetic inference with the same set of data (from known species with consensus phylogenetic relationship), and use two different models for molecular evolution of DNA; one the model we developed, and another conventional model. Comparing the results of the phylogenetic inference with the consensus tree, can give us an insight on the importance of the state of the nucleotide content in phylogenetic inference studies.

Next part of my suggestions is developing a simulation package based on my proposed model of the evolution of the variation of the nucleotide content. This simulation can, indeed, help to understand the magnitude of the importance of each of the causes behind the heterogeneity. This simulation can account for various recombination rates, different mutation frequencies, different selection coefficient values, different proportions of harmful mutations, as

well as different organizations of the chromosomes in terms of genes' lengths, gene densities and some other attributes.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990 Basic Local Alignment Search Tool. *J. Mol. Biol.* 215: 403-410.
- Ballard, J. W. O., and Whitlock, M. C., 2004 The incomplete natural history of mitochondria. *Molecular Ecology* 13: 729–744
- Bentley S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., *et al.*, 2002 Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2). *Nature* 417: 141-147
- Bernardi, G., and Bernardi, G., 1985 Codon Usage and Genome Composition. *J. Mol. Evol.* 22: 363-365
- Bernardi, G., and Bernardi, G., 1986 The Human Genome and Its Evolutionary Context. Cold Spring Harb. Symp. Quant. Biol. 51: 479-487
- Bernardi, G., 1989 The isochore organization of human genome. *Annu. Rev. Genet.* 23: 637-61
- Bernardi, G., 1993. The isochore organization of the human genome and its evolutionary history - a review. *Gene* 135: 57-66
- Bernardi, G., 2000 Isochores and the evolutionary genomics of vertebrates. *Gene* 241: 3-17
- Bernardi, G., 2001 Misunderstanding about isochores. Part 1. *Gene* 276: 3–13
- Birdsell, J.A. 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* 19: 1181–1197
- Blanquart, S., Gascuel, O., 2011 Mitochondrial genes support a common origin of rodent malaria parasites and *Plasmodium falciparum*'s relatives infecting great apes. *BMC Evolutionary Biology* 11:70

- Brown, W. M., George, M., Wilson, A. C., 1979 Rapid evolution of animal mitochondrial DNA. PNAS. 76 (4): 1967-1971
- Brown, W. M., Prager, E. M., Wang, A., Wilson, A. C., 1982 Mitochondrial DNA Sequences of Primates: Tempo and Mode of Evolution. J. Mol. Evol. 18: 225-239
- Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., *et al.*, 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455: 757-763
- Chargaff, E., 1950 Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6: 201-209
- Charlesworth, B., Morgan, M. T., Charlesworth, D., 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303
- Charlesworth, B., 1994 Genetic recombination: patterns in the genome. Curr. Biol. 4:182–184
- Charlesworth, D., Charlesworth, B., Morgan, M. T., 1995 The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619-1632
- Chojnowski, J. L., Franklin, J., Katsu, Y., Iguchi, T., Guillette, Jr. L. J., *et al.*, 2007 Patterns of Vertebrate Isochore Evolution Revealed by Comparison of Expressed Mammalian, Avian, and Crocodilian Genes. J. Mol. Evol. 65: 259-266
- Cohen, N., Dagan, T., Stone, L., Graur, D., 2005 GC Composition of the Human Genome: In Search of Isochores. Mol. Biol. Evol. 22(5): 1260-1272.
- Comings, D. E., Avelino, E., Okada, T. A., Wyandt, H. E., 1973 The mechanism of C- and G-banding of chromosomes. Experimental Cell Research 77: 469-493

- Constantini, M. Clay, O., Auletta, F., Bernardi, G., 2006 An isochore map of human chromosomes. *Genome Research* 16: 536–541
- Comeron, J. M., Kreitman, M., Aguade, M., 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics*. 151(1): 239-49
- Cox, E. C., 1972 On the Organization of Higher Chromosomes. *Nature* 239: 133-134.
- Crozier, R. H., and Crozier, Y. C., 1993 The Mitochondrial Genome of the Honeybee *Apis mellifera*: Complete Sequence and Genome Organization. *Genetics* 133: 97-117
- Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981 The Major Components of the Mouse and Human Genomes. *Eur. J. Biochem.* 115: 227-233
- Dávalos, L. M., Perkins, S. L., 2008 Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. *Genomics* 91: 433-442
- Duret, L., Semon, M., Piganeau, G., 2002 Vanishing GC-Rich Isochores in Mammalian Genomes. *Genetics* 162: 1837-1847.
- Duret, L., and Arndt, P.F. 2008. The impact of recombination on nucleotide substitutions in human genome. *PLoS Genet.* 4: e1000071
- Duret, L., and Mouchiroud, D., 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 96: 4482–4487.
- Echols, H., and Goodman, M. F., 1991 Fidelity Mechanisms in DNA Replication. *Annu. Rev. Biochem.* 60: 477-511.
- Escalante, A. A., and Ayala, F. J., 1994 Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *PNAS* 91: 11373-11377

- Evans A. G., and Wellems, T. E. 2002 Coevolutionary Genetics of Plasmodium Malaria Parasites and Their Human Hosts. *Integ. and comp. Biol.* 42:401-407
- Eyre-Walker, A., 1999 Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. *Genetics* 152: 675-683
- Fitch, W. M., 1967 Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.* 26: 499-507
- Felsenstein, J., 1974 The evolutionary advantage of recombination. *Genetics* 78: 737-756
- Foster, P. G., Jermin L. S., Hickey D. A., 1997 Nucleotide Composition Bias Affects Amino Acid Content in Proteins Coded by Animal Mitochondria. *J. Mol. Evol.* 44:282-288
- Foster, P. G., and Hickey, D. A., 1999 Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetic Reconstructions. *J Mol Evol* 48:284-290
- Frank, A. C., Lobry, J. R., 1999 Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* 238: 65-77
- Frederico, L. A., 1990 A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29(10): 2532-2537
- Fryxell, J. K. and Zuckerkandl, E. 2000 Cytosine Deamination Plays a Primary Role in the Evolution of Mammalian Isochores. *Mol. Biol. Evol.* 17(9): 1371-1383.
- Fullerton S. M. 2001 Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Mol. Biol. Evol.* 18(6): 1139-1142

- Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L., 2001 GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* 159: 907-911
- Galtier, N., 2011 The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC Biology*, 9: 61
- Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., *et al.*, 2002 Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419,: 498-511
- Gu, W., Ju, T., Ma, J., Sun, X., Lu, Z., 2004 The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *BioSystems* 73: 89-97
- Gu, X., Hewett-Emmett, D., Li, W., 1998 Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102/103:383-391
- Gu, X., Li, W., 1998, Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *PNAS* 95(11):5899-5905
- Häring, D., and Kypr, J., 2000 No Isochores in the Human Chromosomes 21 and 22? *Biochemical and Biophysical Research Communications* 280: 567-573
- Hayakawa, T., Culleton, R., Otani, H., Horii, T., Tanabe, K., 2008 Big Bang in the Evolution of Extant Malaria Parasites. *Mol. Biol. Evol.* 25(10):2233-2239
- Hesagawa, M., and Hashimoto, T., 1993 Ribosomal RNA trees misleading? *Nature* 361: 23
- Hill, W. G., and Robertson, A., 1966 The effect of linkage on limits to artificial selection. *Genet. Res.* 8: 269-294
- Jørgensen, F.G., Schierup, M. K., Clark A. G., 2007 Heterogeneity in Regional GC Content and Differential Usage of Codons and Amino Acids in GC-Poor and GC-Rich Regions of the Genome of *Apis mellifera*. *Mol. Biol. Evol.* 24(2): 611-619

- Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in Mammalian Protein Metabolism, edited by H. N. Munro. Academic Press, New York
- Kitazaki K., Kubo, T., 2010 Cost of having the largest mitochondrial genome: evolutionary mechanism of plant mitochondrial genome. J Bot., 620137: 1-12
- Keller, I., Bensasson D., Nichols, R. A., 2007 Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. PLoS Genetics. 3(2): 185-191
- Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., *et al.*, 2009 Ensembl Genomes: Extending Ensembl across the taxonomic space. Nucleic Acids Research Database issue 38: D563-D569
- Kimura, M., 1968 Evolutionary rate at the molecular level. Nature 217: 624-626
- Kimura, M., and Ohta, T., 1974 On Some Principles Governing Molecular Evolution. PNAS 71(7): 2848-2852
- Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120
- King, J. L., and Jukes, T. H., 1969 Non-Darwinian evolution. Science 164: 788-798
- Larkin, M. A., Blackshields, G., Brown, M. P., Chenna, R., McGettigan, P. A., 2007 Clustal W and Clustal X version 2.0. Bioinformatics 23(21): 2947-8
- Lawrence, J. G., Ochman, H., Hartl, D. L., 1991 Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. 137 (8): 1911-1921
- Liao, B. Y., Scott, N. M., Zhang, J., 2006 Impact of Gene Essentiality, Expression Pattern, and Gene Compactness on the evolutionary Rate of Mammalian Proteins. Mol. Biol. Evol. 23 (11): 2072-208

- Lobry, J. R., and Gautier, C., 1994 Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucl. Acids Res.* 22 (15): 3174-3180
- Loomis, W. F., and Smith, D. W., 1990 Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *PNAS* 87: 9093-9097
- Macaya, G., Thiery, J-P., Bernardi, G., 1976 An Approach to the Organization of Eukaryotic Genomes at a Macromolecular level. *J. Mol. Biol.* 108: 237-254
- Marmur, J., and Doty, P., 1959 Heterogeneity in deoxyribonucleic acids. *Nature* 183: 1427-1429
- Meunier, J., and Duret, L., 2004 Recombination Drives the Evolution of GC-Content in the Human Genome. *Mol. Biol. Evol.* 21(6):984-990
- Munoz-Torres, M. C., Reese, J. T., Childers, J. P., Bennett, A. K., Sundaram, J. P., *et al.* 2011 Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Research Database* issue 39: D658–D662
- Muto, A., and Osawa, S., 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. *PNAS* 84: 166-169
- Nabholz, B., Kunstner, A., Wang, R., Jarvis, E. D., Ellegren, H., 2011 Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Mol. Biol. Evol.* 28 (8): 2197-2210
- Nekrutenko, A., and Li, W. H., 2000 Assessment of Compositional Heterogeneity Within and Between Eukaryotic Genomes. *Genome Research* 10:1986-1995
- Nikolaou, C., and Almirantis, Y., 2006 Deviations from Chargaff's second parity rule in organellar DNA insights into the evolution of organellar genomes. *Gene* 381: 34-41

- Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D., Gartl, D. L., 1998 Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572-575
- Pennisi, E., 2001 Malaria's Beginnings: On the Heels of Hoes? *Science* 293(5529): 416-417
- Perkins, S. L., and Schall, J. J., 2002 A molecular phylogeny of malarial parasites recovered from Cytochrome b gene sequences. *J. Parasitol.* 88(5): 972-978
- Rich, S. M., and Ayala, F. J., 2000 Population structure and recent evolution of *Plasmodium falciparum*. *PNAS* 97 (13): 6994-7001
- Ronquist, F., and Huelsenbeck, J. P., 2003, MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19 (12): 1572-1574
- Roure, B., and Philippe, H., 2011 Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology* 11:17
- Rudner, R., Karkas, J. D., Chargaff, E., 1968 Separation of *B. subtilis* DNA into complementary strands, i. biological properties. *PNAS* 60: 630-635
- Saccone, S., De Sario, A., Wiegant, J., Raap, A. K., Della Valle, G., *et al.*, 1993 Correlations between isochores and chromosomal bands in the human genome. *PNAS* 90: 11929-11933
- Sanjuán, R., and Bordería, A. V., 2011 Interplay between RNA Structure and Protein Evolution in HIV-1. *Mol Biol. Evol.* 28 (4): 1333-1338
- Serres-Giardi, L., Belkhir, K., David, J., Glémin, S., 2012 Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *The Plant Cell* 24: 1379-1397
- Shields, D. C., 1990 Switches in species-specific codon preferences: The influence of mutation biases. *J Mol Evol.* 31: 71-80

- Silva, J. C., Egan, A., Friedman, R., Munro, J. B., Carlton, J. M., *et al.*, 2011. Genome sequences reveal divergence times of malaria parasite lineages. *Parasitology* 138(13): 1737-1749
- Singer, G. A. C., and Hickey, D. A., 2000 Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of proteins. *Mol. Biol. Evol.* 17(11): 1581-1588.
- Smith, N. G. C., and Eyre-Walker, A., 2001 Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C-Rich Genes in Humans. *Mol. Biol. Evol.* 18(6): 982-986
- Smith, M. J., and Haigh, J., 1974 The hitch-hiking effect of a favourable gene. *Genetics Research*. 23: 23-35
- Sniegowski, P. D., Gerrish, P. J., Johnson, T., Shaver, A., 2000 The evolution of mutation rates: separating causes from consequences. *BioEssays*. 22: 1057-1066
- Stoletzki, N., 2011 The surprising negative correlation of gene length and optimal codon use—Disentangling translational selection from GC-biased gene conversion in yeast. *BMC Evol. Biol.* 11: 93.
- Sueoka, N., 1959 A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *PNAS* 45(10): 1480–1490
- Sueoka, N., 1961 Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. *Journal of Molecular Biology*. 3(1): 31-40
- Sueoka, N., 1962 On The Genetic Basis Of Variation And Heterogeneity of DNA Base Composition. *PNAS* 48: 582-592.
- Sueoka, N., 1988 Directional mutation pressure and neutral molecular evolution. *PNAS* 85: 2653-2657

- Sueoka, N., 1995 Intrastrand Parity Rules of DNA Base Composition and Usage Biases of Synonymous Codons. *J. Mol. Evol.* 40:318-325
- Tavaré, S., 1986 Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences* (American Mathematical Society) 17: 57-86
- van Passel, M. W. J., 2006 Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* 7:26
- Vinogradov, A. E., 1994. Measurement by flow cytometry of genomic AT_GC ratio and genome size. *Cytometry* 16: 34-40
- Vinogradov, A. E., 2003 DNA helix: the importance of being GC-rich. *Nucleic Acids Research* 31(7): 1838-1844
- Vogel, F., and Rörhborn, G., 1966 Amino-acid Substitutions in Haemoglobins and the Mutation Process. *Nature* 210: 116-117
- Wakeley, J., 1996 The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *TREE* 11(4): 158-162
- Wang, H. C., Singer, G. A. C., Hickey, D. A., 2004 Mutational Bias Affects Protein Evolution in Flowering Plants. *Mol. Biol. Evol.* 21(1): 90-96.
- Watson, J. D., and Crick, F. H. C., 1953 Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171: 737-738
- Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., Oosick, R., 2008 *Molecular Biology of the Gene*. San Francisco: Pearson/Benjamin Cummings.

- Xia, X., Wei, T., Xie, Z., Danchin, A., 2002. Genomic Changes in Nucleotide and Dinucleotide Frequencies in *Pasteurella multocida* Cultured Under High Temperature. *Genetics* 161:1385-1394
- Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution* 26(1): 1-7
- Xia, X., Lemey, P., 2009. Assessing substitution saturation with DAMBE. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Cambridge University Press: 611-626
- Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39: 306-314
- Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analysis. *TREE* vol. 11(9): 367-372

Appendices

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Ter	TGA	Ter	A
	TTG	Leu	TCG	Ser	TAG	Ter	TGG	Trp	G

C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G

A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G

G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Figure S.1. Codon tables. (A): Standard codon table

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Ter	TGA	Trp	A
	TTG	Leu	TCG	Ser	TAG	Ter	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Met	ACA	Thr	AAA	Lys	AGA	Ter	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Ter	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Figure S.1. Codon tables. (B): Vertebrate mitochondrial codon table

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Ter	TGA	Trp	A
	TTG	Leu	TCG	Ser	TAG	Ter	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Met	ACA	Thr	AAA	Lys	AGA	Ser	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Ser	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Figure S.1. Codon tables. (C): Invertebrate mitochondrial codon table

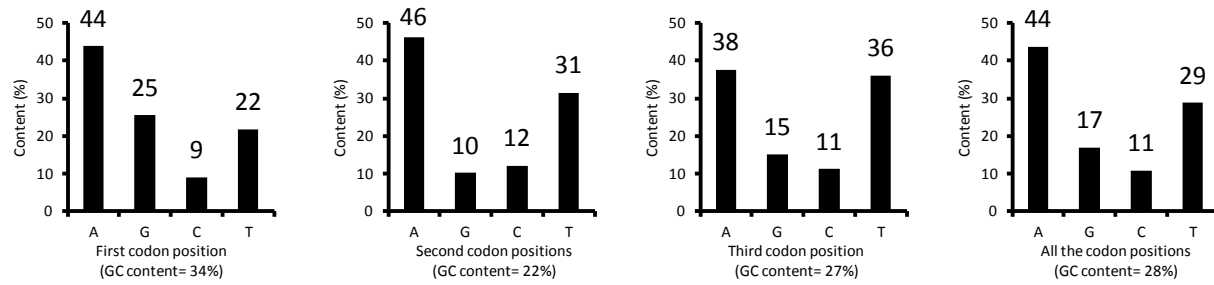


Figure S.2. Individual nucleotide content of the dataset of orthologous coding sequences in *P. falciparum* and *P. vivax*. **(A)** These histograms represent the individual nucleotide content of only conserved sites between the two orthologous subsets of coding sequences in different codon positions.

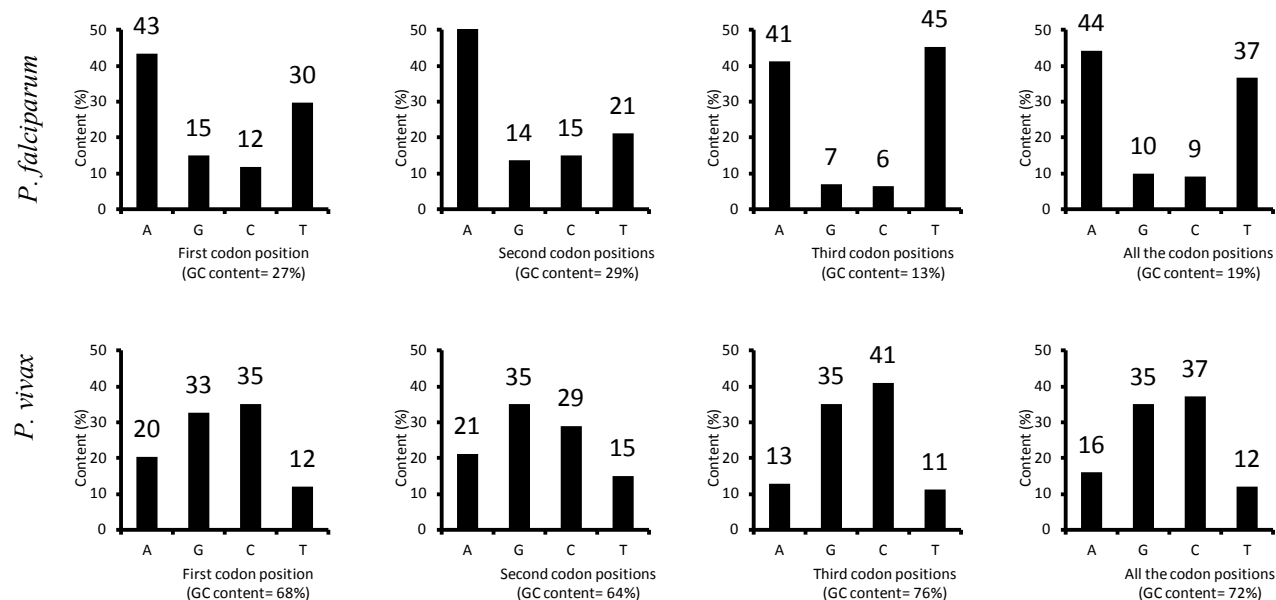


Figure S.2. Individual nucleotide content of the dataset of orthologous coding sequences in *P. falciparum* and *P. vivax*. **(B)** These histograms represent the GC content of only variable sites between the two orthologous subsets of coding sequences in different codon positions. The top row of histograms represents the data for *P. falciparum* and the white histograms in the bottom row represent the data for *P. vivax*.

Figure S.3. Nucleotide substitution matrices for three concatenated, aligned orthologous mitochondrial genes in *P. falciparum* and *P. vivax*.

All the data in Tables (C) to (H) are represented in actual numbers.

C)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	311	10	1	17
	G	17	205	0	7
	C	4	0	107	17
	T	10	4	13	380

Table (C) contains the data for all sites in first codon position.

D)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A		10	1	17
	G	17		0	7
	C	4	0		17
	T	10	4	13	

Table (D) contains only the data for the variable sites in first codon position.

E)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	210	2	2	1
	G	2	161	2	2
	C	2	4	217	5
	T	4	0	9	480

Table (E) contains the data for all sites in second codon position.

F)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A		2	2	1
	G	2		2	2
	C	2	4		5
	T	4	0	9	

Table (F) contains only the data for the variable sites in second codon position.

G)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	370	5	3	104
	G	13	50	1	3
	C	5	0	17	50
	T	97	1	25	359

Table (G) contains the data for all sites in third codon position.

H)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A		5	3	104
	G	13		1	3
	C	5	0		50
	T	97	1	25	

Table (H) contains only the data for the variable sites in third codon position.

Figure S.4. Nucleotide substitution matrices for five concatenated, aligned orthologous genes in *Human* and *Chimpanzee*.

All the data in Tables (A) and (B) are represented in actual numbers.

A)

		Chimpanzee			
		A	G	C	T
Human	A	4448	46	19	19
	G	93	5045	24	30
	C	48	39	5100	127
	T	29	13	28	3630

Table (A) contains the data for all sites (conserved and variable sites together) in all codon positions.

B)

		Chimpanzee			
		A	G	C	T
Human	A		46	19	19
	G	93		24	30
	C	48	39		127
	T	29	13	28	

Table (B) contains the data for only variable sites in all codon positions.

Figure S.5. Nucleotide substitution matrices for 20 concatenated, aligned orthologous genes in *P. falciparum* and *P. vivax*.

All the data in Tables (A) and (B) are represented in actual numbers.

A)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	13711	6271	3583	1829
	G	1229	5175	973	444
	C	768	810	3087	614
	T	1885	2227	5147	8755

Table (A) contains the data for all sites (conserved and variable sites together) in all codon positions.

B)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A		6271	3583	1829
	G	1229		973	444
	C	768	810		614
	T	1885	2227	5147	

Table (B) contains the data for only variable sites in all codon positions.

Figure S.6. Nucleotide substitution matrices for six concatenated, aligned orthologous genes in *P. falciparum* and *P. vivax*

All the data in Tables (A) and (B) are represented in actual numbers.

A)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A	6115	2209	1196	509
	G	279	2298	257	104
	C	207	215	1411	196
	T	448	659	2191	4056

Table (A) contains the data for all sites (conserved and variable sites together) in all codon positions.

B)

		<i>P. vivax</i>			
		A	G	C	T
<i>P. falciparum</i>	A		2209	1196	509
	G	279		257	104
	C	207	215		196
	T	448	659	2191	

Table (B) contains the data for only variable sites in all codon positions.

Figure S.7. Nucleotide substitution matrices for three concatenated, aligned orthologous mitochondrial genes in *P. falciparum* and *P. reichenowi*

All the data in Tables (A) and (B) are represented in actual numbers

A)

		<i>P. reichenowi</i>			
		A	G	C	T
<i>P. falciparum</i>	A	1003	10	8	16
	G	7	454	0	2
	C	5	0	399	24
	T	16	2	20	1346

Table (A) contains the data for all sites (conserved and variable sites together) in all codon positions.

B)

		<i>P. reichenowi</i>			
		A	G	C	T
<i>P. falciparum</i>	A		10	8	16
	G	7		0	2
	C	5	0		24
	T	16	2	20	

Table (B) contains only the data for the variable sites in all codon positions.

Figure S.8. Nucleotide substitution matrices for three concatenated, aligned orthologous mitochondrial genes in *Human* and *Chimpanzee*

All the data in Tables (A) and (B) are represented in actual numbers.

A)

		Chimpanzee			
		A	G	C	T
Human	A	901	45	6	2
	G	44	456	0	1
	C	4	2	978	117
	T	4	0	100	802

Table (A) contains the data for all sites (conserved and variable sites together) in all codon positions

B)

		Chimpanzee			
		A	G	C	T
Human	A		45	6	2
	G	44		0	1
	C	4	2		117
	T	4	0	100	

Table (B) contains only the data for the variable sites in all codon positions.

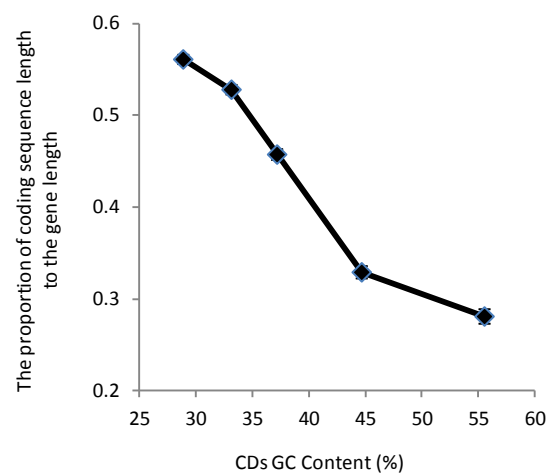


Figure S.9. The correlation between the proportion of coding sequences in total gene length and GC content in five groups of different average GC content.

Table S.1. Nucleotide content in different sites of five genes in *Human* and *Chimpanzee*

	A		G		C		T	
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp
Total Percentage	24	25	28	27	28	28	20	20
Conserved Percentage	24	24	28	28	28	28	20	20
Variable Percentage	16	33	29	19	42	14	14	34

The top row represents the data in all sites together, the conserved sites only is shown in the middle, and the bottom row represents the data for variable sites only. The data are shown in percentages. The three concatenated genes include are the same as the previous tables.

Table S.2. Nucleotide content of five genes in *Human* and *Chimpanzee*

	A		G		C		T		Alignment length
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Gene 1	140	138	146	155	154	146	136	137	576
Gene 2	393	391	351	353	280	278	293	295	1317
Gene 3	3041	3029	3513	3519	3765	3767	2476	2480	12795
Gene 4	206	206	333	334	334	332	201	202	1074
Gene 5	411	412	445	444	353	353	333	333	1542
Average	838	835	958	961	977	975	688	689	3461
99% C.I.	4763	4744	5516	5521	6008	6017	3861	3866	20139

The data shown above represents the actual number of different nucleotides in all the sites (conserved and variable sites together). GI numbers for genes 1 to 5 for human are: 62896524, 37067, 197927451, gb_BC015350.1, 19068220 and for Chimpanzee: 332808314, 345110609, 332816968, 332853013, and 332809722.

Table S.3. Nucleotide content of five genes in *Human* and *Chimpanzee*

	A		G		C		T		GC (%)		GC Difference (%)
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Gene 1	24	24	25	27	27	25	24	24	52	52	0
Gene 2	30	30	27	27	21	21	22	22	48	48	0
Gene 3	24	24	27	28	29	29	19	19	57	57	0
Gene 4	19	19	31	31	31	31	19	19	62	62	0
Gene 5	27	27	29	29	23	23	22	22	52	52	0
Average	25	25	28	28	26	26	21	21	54	54	0
99% C.I.	15	15	8	7	16	16	8	8	21	21	0

The data shown above represent the nucleotide content in percentages in all the sites (conserved and variable sites together). GI numbers are the same as the previous tables.

Table S.4. Nucleotide content of the variable sites in five genes in *Human* and *Chimpanzee*

	A		G		C		T		Number of variable sites
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Gene 1	20	18	16	25	21	13	13	14	70
Gene 2	4	2	2	4	4	2	2	4	12
Gene 3	29	17	17	23	23	25	15	19	84
Gene 4	1	1	0	1	2	0	0	1	3
Gene 5	0	1	1	0	1	1	1	1	3
Average	11	8	7	11	10	8	6	8	34
99% C.I.	50	34	33	48	42	41	28	32	152

The data shown above represents the actual number of different nucleotides in variable sites only. GI numbers are the same as the previous tables.

Table S.5. Nucleotide content of the variable sites in five genes in *Human* and *Chimpanzee*

	A		G		C		T		GC (%)		GC Difference (%)
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Gene 1	29	26	23	36	30	19	19	20	53	54	1
Gene 2	33	17	17	33	33	17	17	33	50	50	0
Gene 3	35	20	20	27	27	30	18	23	48	57	10
Gene 4	33	33	0	33	67	0	0	33	67	33	-33
Gene 5	0	33	33	0	33	33	33	33	67	33	-33
Average	26	26	19	26	38	20	17	29	57	46	-11
99% C.I.	57	29	47	57	62	50	46	26	36	44	79

The data shown above represents the actual number of different nucleotides in variable sites only. GI numbers are the same as the previous tables.

Table S.6. Nucleotide content in different sites of three mitochondrial genes in *P. falciparum* and *P. reichenowi*

	A		G		C		T	
	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>
Total Percentage	31	31	14	14	13	13	42	42
Conserved Percentage	31	31	14	14	12	12	42	42
Variable Percentage	31	25	8	11	26	25	35	38

The top row represents the data in all sites together, the conserved sites only is shown in the middle, and the bottom row represents the data for variable sites only. The data are shown in percentages. The three concatenated genes include Cox I, Cox III, and Cyt b.

Table S.7. Nucleotide content of three mitochondrial genes in *P. falciparum* and *P. reichenowi*

	A		G		C		T		Alignment length
	P. falciparum	P.reich	P. falciparum	P.reich	P. falciparum	P.reich	P. falciparum	P.reich	
Cox I	436	440	219	218	185	189	594	587	1434
Cox III	248	238	84	89	91	90	327	333	750
Cyt b	353	353	160	159	152	148	463	468	1128
Average	346	344	154	155	143	142	461	463	1104
99% C.I.	540	580	388	370	273	285	765	728	1962

The data shown above represents the actual number of different nucleotides in all the sites (conserved and variable sites together).

Table S.8. Nucleotide content of three mitochondrial genes in *P. falciparum* and *P. reichenowi*

	A		G		C		T		GC (%)		GC diff. (%)
	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	
Cox I	30	31	15	15	13	13	41	41	28	28	0
Cox III	33	32	11	12	12	12	44	44	23	24	0
Cyt b	31	31	14	14	13	13	41	41	28	27	1
Average	32	31	14	14	13	13	42	42	26	26	0
99% C.I.	8	3	12	10	4	4	8	11	15	13	3

The data shown above represent the nucleotide content in percentages in all the sites (conserved and variable sites together).

Table S.9. Nucleotide content of the variable sites in three mitochondrial genes in *P. falciparum* and *P. reichenowi*

	A		G		C		T		Number of variable sites
	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	<i>P. falciparum</i>	<i>P. reich</i>	
Cox I	14	10	4	5	13	9	13	20	44
Cox III	6	16	7	2	10	11	10	10	39
Cyt b	8	8	1	2	5	9	5	8	27
Average	9	11	4	3	9	10	9	13	37
99% C.I.	24	24	17	10	23	7	23	37	50

The data shown above represents the actual number of different nucleotides in variable sites only.

Table S.10. Nucleotide content of the variable sites in three mitochondrial genes in *P. falciparum* and *P. reichenowi*

	A		G		C		T		GC (%)		GC diff. (%)
	<i>P. falciparum</i>	<i>P.reich</i>	<i>P. falciparum</i>	<i>P.reich</i>	<i>P. falciparum</i>	<i>P.reich</i>	<i>P. falciparum</i>	<i>P.reich</i>	<i>P. falciparum</i>	<i>P.reich</i>	
Cox I	32	23	9	11	30	20	30	45	39	32	7
Cox III	15	41	18	5	26	28	41	26	44	33	10
Cyt b	30	30	4	7	19	33	48	30	22	41	-19
Average	26	31	10	8	25	27	40	34	35	35	0
99% C.I.	51	53	41	18	32	37	54	60	64	27	90

The data shown above represent the nucleotide content in percentages in variable sites only.

Table S.11. Nucleotide content in different sites of three mitochondrial genes in *Human* and *Chimpanzee*

	A		G		C		T	
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp
Total Percentage	28	28	14	15	32	31	26	27
Conserved Percentage	29	29	15	15	31	31	26	26
Variable Percentage	16	16	14	14	38	33	32	37

The top row represents the data in all sites together, the conserved sites only is shown in the middle, and the bottom row represents the data for variable sites only. The data are shown in percentages. The three concatenated genes include Cox I, Cox III, and Cyt b.

Table S.12. Nucleotide content of three mitochondrial genes in *Human* and *Chimpanzee*

	A		G		C		T		Alignment length
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Cox I	417	409	249	256	463	451	410	423	1539
Cox III	210	214	116	112	249	245	208	212	783
Cyt b	327	330	136	135	389	388	288	287	1140
Average	318	318	167	168	367	361	302	307	1154
99% C.I.	594	1386	411	1093	622	1491	583	1511	2166

The data shown above represents the actual number of different nucleotides in all the sites (conserved and variable sites together).

Table S.13. Nucleotide content of three mitochondrial genes in *Human* and *Chimpanzee*

	A		G		C		T		GC (%)		GC diff. (%)
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Cox I	27	27	16	17	30	29	27	27	46	46	0
Cox III	27	27	15	14	32	31	27	27	47	46	-1
Cyt b	29	29	12	12	34	34	25	25	46	46	0
Average	28	28	14	14	32	32	26	27	46	46	0
99% C.I.	6	17	12	34	12	34	4	17	2	3	0

The data shown above represent the nucleotide content in percentages in all the sites (conserved and variable sites together).

Table S.14. Nucleotide content of the variable sites in three mitochondrial genes in *Human* and *Chimpanzee*

	A		G		C		T		Number of variable sites
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Cox I	26	18	16	23	51	39	36	49	129
Cox III	10	14	12	8	26	22	22	26	70
Cyt b	17	20	17	16	46	45	46	45	126
Average	18	17	15	16	41	35	35	40	108
99% C.I.	113	43	37	106	187	169	170	174	606

The data shown above represents the actual number of different nucleotides in variable sites only.

Table S.15. Nucleotide content of the variable sites in three mitochondrial genes in *Human* and *Chimpanzee*

	A		G		C		T		GC (%)		GC diff. (%)
	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	Human	Chimp	
Cox I	20	14	12	18	40	30	28	38	52	48	-4
Cox III	14	20	17	11	37	31	31	37	54	43	-11
Cyt b	13	16	13	13	37	36	37	36	50	48	-2
Average	16	17	14	14	38	32	32	37	52	46	-6
99% C.I.	51	44	35	48	23	41	61	16	30	44	73

The data shown above represent the nucleotide content in percentages in variable sites only.